UNIVERSIDADE TÉCNICA DE LISBOA

Faculdade de Medicina Veterinária

Structural and functional insights into the role of Carbohydrate Esterases and Carbohydrate-Binding Modules in plant cell wall hydrolysis

Márcia Alexandra da Silva Correia

CONSTITUIÇÃO DO JÚRI

PRESIDENTE

Reitor da Universidade Técnica de Lisboa

VOGAIS

Doutora Maria dos Anjos Lopez de Macedo

Doutora Maria Joao Lobo de Reis Madeira Crispim Romao

Doutor Shabir Najmudim

Doutor José António Mestre Prates

Doutor Carlos Mendes Godinho de Andrade Fontes

ORIENTADOR

Doutor José António Mestre Prates

CO-ORIENTADOR

Doutor Carlos Mendes Godinho de Andrade Fontes

2009

LISBOA

UNIVERSIDADE TÉCNICA DE LISBOA

Faculdade de Medicina Veterinária

Structural and functional insights into the role of Carbohydrate Esterases and Carbohydrate-Binding Modules in plant cell wall hydrolysis

Márcia Alexandra da Silva Correia

Tese de Doutoramento em Ciências e Tecnologia Animal

CONSTITUIÇÃO DO JÚRI

PRESIDENTE

Reitor da Universidade Técnica de Lisboa

VOGAIS

Doutora Maria dos Anjos Lopez de Macedo

Doutora Maria Joao Lobo de Reis Madeira Crispim Romao

Doutor Shabir Najmudim

Doutor José António Mestre Prates

Doutor Carlos Mendes Godinho de Andrade Fontes

ORIENTADOR

Doutor José António Mestre Prates

CO-ORIENTADOR

Doutor Carlos Mendes Godinho de Andrade Fontes

2009

LISBOA

Ao Henrique e à minha pequena Inês

Aos meus pais, Maria Helena e José, e à minha irmã Joana

ACKNOWLEDGEMENTS / AGRADECIMENTOS

À Faculdade de Medicina Veterinária da Universidade Técnica de Lisboa por me ter proporcionado meios humanos e materiais para a execução da maior parte do trabalho que originou esta tese.

Ao CIISA – Centro de Investigação Interdisciplinar em Sanidade Animal (FMV-UTL), na pessoa do seu coordenador, Professor Doutor Luís Tavares, pelos meios técnicos disponibilizados para a realização deste projectos, bem como no apoio e na divulgação em vários congressos dos resultados obtidos neste trabalho.

To the Institute for Cell and Molecular Biosciences from the Newcastle University, for have given me the human and material means for accomplishing part of this thesis and for the wellcoming way they have received me.

Ao Professor Doutor José Mestre Prates, meu orientador ciêntifico, pela forma como me integrou no Laboratório de Bioquímica, pela total disponibilidade, pelo esclarecimento de todas as duvidas e pela sua extrema organização.

Ao Professor Doutor Carlos Fontes, meu co-orientador cientifico, por ter estado sempre presente, pela sua imensa capacidade de trabalho, por todo o seu empenho, esforço e conselhos preciosos ao longo de todo este projecto, por me ensinar o caminho a percorrer, e finalmente, não menos importante, pela boa disposição presente ao longo deste anos. Um muito obrigada.

To Professor Harry Gilbert, from Newcastle University Upon Tyne, UK, for received me so warmly in his laboratory and for all is knowledge and assistance given when i was working in Newcastle.

To Cedric Montaneir for all the help during my stay in Newcastle and for his contribution to the Rhe-CBM35 and *Ct*CBM55 study.

Ao prof. Doutor Luis Ferreira pelos conselhos, bom humor e pela forma hospitaleira como me recebeu;

À Professora Doutora Maria João Romão, e à Doutora Ana Luisa Carvalho, da Faculdade de Ciência e Tecnologia da Universidade Nova de Lisboa, pela partilha de meios técnico e de conhecimento ao longo deste Doutoramento.

Aos meus amigos e colegas de laboratório, à Benedita, à Catarina, à Cristina, ao Fernando, à Helena, à Joana, à Marija, à Marisa, à Patricia, à Paula, à Susana, à Teresa, à virginia, à Vânia e à Zé, pela boa disposição, pelo incentivo, por toda a ajuda e pela amizade demonstradra em momentos únicos da minha vida. À Benedita, um obrigada especial por me acompanhar juntamente com a sua amizade desde os inicios da faculdade. À Helena, para além da amizade e simpatia, pela grande técnica que ela é, sempre presente com a sua ajuda.

Aos meus amigos extra-trabalho, que apesar de nem sempre estarmos juntos a quantidade de vezes desejada, estiveram sempre presentes, com alegria, divertimento e apoio, nos momentos mais marcantes da minha vida.

A toda a minha familia, que esteve sempre presente em todos os momentos felizes, calmos e stressantes deste percuso, por cada um deles ser como é.

Aos meus pais, pelos seus valores, pelo seu amor, por terem contribuido para eu ser a pessoa que sou hoje e por me acompanharem em todas a fases desta grande caminhada que é a Vida. Adoro-vos.

À minha irmã Joana, pelo seu afecto, pelo seu orgulho e confiança que sempre depositou em mim.

Ao Henrique, à minha luz, por ser um marido incansável, que me apoiou em todos os momentos cheio de orgulho e de Amor. Amo-te muito.

À minha Croquetita, à minha Estrela, à minha Inês, que esteve inevitavelmente comigo em todas as ocasiões e que me deu inspiração e força para finalizar a escrita deste trabalho.

Por último gostaria de estender os meus agradecimentos a todos aqueles que de uma forma ou de outra (fornecendo ideias e/ou criticando), foram ajudando nas inúmeras etapas ao longo desta caminhada.

A todos, os meus sinceros agradecimentos.

This work was funded by Fundação para a Ciência e Tecnologia through individual fellowship SFRH/BD/23784/2005.
Co-funded by POCTI/CVT/2004/61162, POCTI/BIA-PRO/2004/59118 and FSE from Ministério da Ciência, Tecnologia e Ensino Superior.

ABSTRACT

Structural and functional insights into the role of Carbohydrate Esterases and Carbohydrate-Binding Modules in plant cell wall hydrolysis

Plant cell wall polysaccharides offer an extraordinary source of carbon and energy that can be used by various microorganisms, thus constituting a central component of the carbon cycle. The anaerobic thermophilic bacterium *Clostridium thermocellum* is one of the most prolific degraders of plant cell wall polysaccharides. It produces a multi-enzyme extra-cellular complex of cellulases and hemicellulases, the cellulosome, and these enzymes were shown to have a remarkable biotechnological potential. Based on the recently determined genome sequence of *Clostridium thermocellum*, we aimed to address several unresolved questions concerning the mechanism of plant cell wall hydrolysis by microbial multi-enzyme complexes.The crystal structure and biochemical properties of the N-terminal carbohydrate esterase domain of *Clostridium thermocellum* *Ct*Ces3-1 were determined (chapter 2). The enzyme is a thermostable acetyl-specific esterase that exhibits a strong preference for acetylated xylan. In adition, we report, in Chapter 3, the characterization of four carbohydrate-binding modules (CBMs) of family 35 that display specificity for Δ4,5-anhydrogalacturonic acid (Δ4,5-GalA), although two of the proteins also interact with glucuronic acid (GlcA). X-ray crystallographic data revealed that the ligand binding site is highly conserved in the four CBM35s. In chapter 4, biochemical properties of a CBM6 (CBM6-1) from *Cellvibrio mixtus* *Cm*Cel5A are presented. The data revealed that CBM6-1 recognizes specifically β1,3-glucans through a previous unknown ligand binding plataform. These studies reveal the different mechanisms by which a highly conserved protein platform (CBM6) can recognise a common ligand. Finally, we identified a novel CBM within the large *C. thermocellum* cellulosomal protein Cthe_2193 (*Ct*CBM55), which is the founder member of a new CBM family. *Ct*CBM55, in contrast to the previously characterized cellulosomal CBMs, binds to D-galactose and L-arabinose in either anomeric configuration in complex polysaccharides. Ligand specificity is conferred through numerous interactions with the axial *O4* of the target sugars, a feature that distinguishes galactose and arabinose from the other major sugars located into plant cell walls.

Key-words: *Clostridium thermocellum,* carbohydrate esterases, carbohydrate-binding modules.

RESUMO

Perspectivas Estruturais e funcionais do papel das carbohidrato esterases e dos modulos de ligação a hidratos de carbono na hidrólise da parede cellular vegetal.

Os polissacáridos da parede celular vegetal constituem uma fonte de carbono e energia que pode ser utilizada por diversos microorganismos desempenhando assim um papel de relevo no ciclo do carbono. A bactéria termofílica anaeróbia *Clostridium thermocellum* é muito eficaz na degradação dos polissacáridos da parede celular das plantas. Esta produz um complexo multi-enzimático extra-celular de celulases e hemicelulases, denominado celulossoma, demonstrando estas enzimas um extraordinário potencial biotecnológico. Neste trabalho foram determinadas as propriedades bioquímicas e a estrutura cristalográfica do domínio catalítico N-terminal da enzima *Ct*Ces3 (*Ct*Ces3-1) do *Clostridium thermocellum* (Capítulo 2). Esta enzima é uma esterase termostável específica para grupos acetilo e que demonstra uma forte preferência por xilano acetilado. É também feita, no Capitulo 3, a caracterização de quatro módulos de ligação a hidratos de carbono (CBMs) da familia 35 que demonstram afinidade para o ácido Δ4,5-anidrogalacturónico (Δ4,5-GalA), apesar de duas das proteínas também interagirem com o acido glucurónico (GlcA). Os dados cristalográficos demonstram que o local de ligação é altamente conservado nos quatro CBM35s. No Capítulo 4, são reveladas as propriedades bioquímicas de um CBM6 da enzima *Cm*Cel5A (CBM6-1) do *Cellvibrio mixtus*. Os dados revelaram que o CBM6-1 reconhece especificamente β1,3-glucanos através de uma plataforma de ligação previamente desconhecida. Estes estudos demonstram os diferentes mecanismos pelos quais uma plataforma proteíca altamente conservada (CBM6) pode reconhecer o mesmo ligando. Finalmente, identificámos um novo CBM da enzima Cthe_2193 (*Ct*CBM55), pertencente ao celulossoma do *C. thermocellum*, estabelecendo-se assim uma nova família de CBMs. O *Ct*CBM55, ao contrário de outros CBMs do celulossoma, liga-se à D-galactose e à L-arabinose em qualquer configuração anomérica em polissacáridos complexos. A sua especificidade é conferida por varias interacções com o *O4* axial dos açucares alvo, uma característica que distingue a galactose e a arabinose dos outros hidratos de carbono que compoêm a parede celular vegetal.

Palavras-Chave: *Clostridium thermocellum,* carbohidrato esterases, módulos de ligação a hidratos de carbono.

PAPERS

This thesis is based on the following publications:

Correia, M. A. S., Prates, J. A. M., Brás, J., Fontes, C. M. G. A., Newman, J. A., Lewis, R. J., Gilbert, H. J. and Flint, J. E. (2008). Crystal Structure of a Cellulosomal Family 3 Carbohydrate Esterase from Clostridium thermocellum Provides Insights into the Mechanism of Substrate Recognition. *Journal of Molecular Biology*. **379**, 64-72.

Montanier, C., Bueren, A. L. v., Dumon, C., Flint, J. E., **Correia, M. A.**, Prates, J. A., Firbank, S. J., Lewis, R. J., Grondin, G. G., Ghinet, M. G., Gloster, T. M., Cecile Hervef, Knox, J. P., Talbot, B. G., Turkenburg, J. P., Kerovuo, J., Brzezinski, R., Fontes, C. M. G. A., Davies, G. J., Boraston, A. B. and Gilbert, H. J. (2009). Evidence that family 35 carbohydrate-binding modules display conserved specificity but divergent function. *PNAS*, **106** (9), 3065–3070.

Correia, M. A. S., Pires, V. M. R., Gilbert, H. J., Bolam, D. N., Prates, J. A. M., Alves, V. D., Ferreira, L. M. A. and Fontes, C. M. G. A. (2009). Family 6 carbohydrate-binding modules display multiple β-1,3-linked glucan specific binding interfaces. Submitted for publication.

Correia, M. A. S., Montanier, C., Flint, J. E., Prates, J. A. M., Faribanks, S., Lewis, R. J., Coutinho, P. M., Fontes, C. M. G. A. and Gilbert, H. J. (2009). A novel non-catalytic carbohydrate-binding modules displays specificity for galactose-containing polysaccharides. Submitted for publication.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS AND SYMBOLS

A	Alanine
Å	Angstrom
AGE	Affinity gel electrophoresis
*Bh*CBM6	CBM6 from *Bacillus haludurans*
BSA	Bovine serum albumine
°C	Degree Celsius
Cα	Alpha carbon
CAZy	Carbohydrate-active enzyme
CBD	Cellulose-binding domain
CBM	Carbohydrate-binding module
CBM6	Carbohydrate-binding module family 6
CBM22	Carbohydrate-binding module family 22
CBM35	Carbohydrate-binding module family 35
CCP4	Collaborative computational project number 4
CE	Carbohydrate esterase
CE3	Carbohydrate esterase family 3
Chi-CBM35	CBM35 of the *exo*-β-D-glucosaminidase CsxA from *Amycolatopsis orientalis*
CMC	Carboxymethylcellulose
CsxA	exo-β-D-glucosaminidase from *Amycolatopsis orientalis*
Csn	endo-chitosanase from *Amycolatopsis orientalis*
Ct	*Clostridium thermocellum*
*Ct*CBM55	*Clostridium thermocellum* CBM55
CtCes3	*Clostridium thermocellum* CE3
Δ4,5-GalA	Δ 4,5-anhydrogalaturonic acid
Da	Dalton
DP	Degree of polymerization
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
E	Glutamic acid / Glutamate
EC	Enzyme comission

EDTA	Ethylenediamine tetraacetic acid
ESRF	European synchrotron radiation facility
Fe	Iron
FPLC	Fast Protein Liquid Chromatography
Fuc	Fucose
Gal	Galactose
Gal_2Man_5	6^3, 6^4-di-α-D-galactosyl mannopentaose
$GalMan_3$	6^1-α-D-galactosyl mannotriose
GAX	Glucuronoarabinoxylan
GH	Glycoside hydrolase
GH5	Glycoside hydrolase family 5
GH10	Glycoside hydrolase family 10
GH11	Glycoside hydrolase family 11
GlcA	Glucuronic acid
Gln	Glutamine
Glu	Glutamic acid / Glutamate
HEC	Hydroxyethylcellulose
Hepes	4-(2-Hydroxyethyl)-1-piperazine-ethanesulfonic acid
HG	Homogalacturonan
HPLC	High-performance liquid chromatography
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl 1-thio-β-d-galactopyranoside
ITC	Isothermal titration calorimetry
IU	International units
IUB-MB	International Union of Biochemistry and Molecular Biology
K_a	Equilibrium affinity constant
k_{cat}	Catalytic efficiency
kDa	KiloDalton
K_m	Michaelis constant
kW	Kilowatt
LB	Luria-Bertani groth medium
LK	Linker
M	Molar (mol L^{-1})

MAD	Multi-wavelengh anomalous dispersion
mA	Miliampere
MES	4-Morpholine-ethanesulfonic acid
NMR	Nuclear Magnetic Resonance spectroscopy
°	Degree
OD	Optical density
PCR	Polimerase Chain Reaction
PDB	Protein data bank
PEG	Polyethyleneglycol
Pel-CBM35	CBM35 of a pectate lyase
pH	Potential of hidrogen
PL	Polysaccharide Lyases
PNPAc	4-ρ-nitrophenyl acetate
R	Arginine
Rhe-CBM35	CBM35 of the rhamnogalauronan acetyl esterase from *Clostridium thermocellum*
RGI	Rhamnogalacturonan I
RGII	rhamnogalacturonan II
rpm	Rotation per minute
SAD	single-wavelength anomalous dispersion
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SeMet	Seleno-methionine
SLH	S-layer homology module
Tris	2-Amino-2-hydroxymethyl-1,3-propanediol
U	Units
V	Volt
X	Glucose decorated α-1,6 with xylose
XG	Xyloglucan
XGOs	Xylogluco-oligosaccharides
X-gal	5-bromo-4-chloro-3-indolyl β-D-galactoside
Xyl	A-1,6 D-xylosyl
Xyl-CBM35	CBM35 of a xylanase from *Cellvibrio japonicus*
%	percent

INTRODUTION

Plant cells are encased within a complex polysaccharide rich cell wall, which constitutes the raw material that is used to manufacture textiles, paper, lumber, films, thickeners and other products. The plant cell wall is also the primary source of cellulose, one of the most abundant and useful biopolymers on Earth. The biological degradation of plant cell wall polysaccharides involves the cooperative action of a large range of Glycoside Hydrolases (GHs), Carbohydrtae Esterases (CEs) and Polysaccharide Lyases (PLs), the majority of these being secreted by microorganisms like bacteria and fungi. Enzymes that digest plant cell wall polysaccharides generally contain non-catalytic domains, carbohydrate-binding domains (CBMs), that function by anchoring the enzyme to the substrate, thus potentiating catalytic activity (Tunnicliffe et al., 2005).

A generic feature of enzymes that hydrolyze complex carbohydrates is their modular structure. The most common noncatalytic modules are the carbohydrate-binding modules (CBMs), which are grouped into sequence-based families (Boraston *et al*., 2004). The general function of CBMs is to promote the interaction of the enzyme with the target substrate, thereby increasing the efficiency of catalysis (Boraston *et al*., 2003; Hall *et al.*, 1995; Tomme *et al.*¸1988). CBMs are prevalent in plant cell wall degrading enzymes where they direct the cognate catalytic modules to their target substrate within these complex composite structures (Boraston *et al*., 2004). CBMs are the perfect candidates for a range of biotechnological applications since they present three basic properties that were described by Shoseyov *et al.* (2006): (i) CBMs are usually independently folding units and, therefore, can function autonomously in chimeric proteins; (ii) the attachment matrices are abundant and inexpensive and also have excellent chemical and physical properties; and (iii) the binding specificities can be controlled, and therefore the right solution can be adapted to an existing problem (Shoseyov *et al.* 2006).

Large-scale recovery and purification of biologically active molecules continues to be a challenge for many biotechnology companies. Various purification procedures have been developed, of which biospecific affinity purification (affinity chromatography) has become one of the most rapidly developing divisions of immobilized affinity ligand technology. To date, several affinity tags have been developed that vary in size from several amino acids to a complete protein and each

individual affinity-based purification system embodies specific advantages. When compared with most immobilization systems, cellulose is an economical support-matrix for large-scale protein purification (Levy and Shoseyov, 2002). CBMs are high-capacity purification tags that may be used for the isolation of biologically active target peptides and many protein entities at relatively low cost, (Shoseyov *et al.* 2006). Protein engineering using CBMs is also an emerging field and it is well established that expression of foreign proteins fused to CBMs results, for the most part, in high expression levels, and as a result, expression vectors (pET34 to pET38 from Novagen) incorporating CBMs as fusion tags were developed. Several studies have shown the potential of CBMs for modifying the characteristics of several enzymes and therefore, the basic approach in CBM engineering was to replace or add a CBM in order to improve hydrolytic activity (Shoseyov *et al.* 2006).

Another area of increasing interest for CBMs is the production of bioethanol from cellulosic material. Lignocellulose is the most abundant renewable natural resource on earth. CBMs are pivotal for the conversion of this organic substrate into fuels since they are able to target the catalytic modules of polysaccharidases to their target bonds. Resulting sugars may then be converted to liquid fuel (Shoseyov *et al.* 2006). The strong affinity that exists between cellulose and CBMs is also used in many applications associated with the textile industry. Numerous laundry powders contain recombinant enzymes that do not possess a native affinity to the cellulosic fabric (amylases, proteases, lipases, and oxidoreductases). The performance of these enzymes, under conventional washing conditions, can be improved by increasing their affinity to the textile substrate by fusion to CBMs (Levy and Shoseyov, 2002). Additional substances can also be targeted to cellulosic fabrics through the use of CBMs, as for instant fragrance-bearing particles in the case of laundry powder. It was also discovered that applying CBMs to cellulose fibers has a potential for use in paper recycling since it has been demonstrated that the application of CBMs on secondary fibers, such as old paperboard containers, results in increased tensile and burst indexes as well as improvement in pulp drainage (Levy and Shoseyov, 2002).

Additional biotechnological uses for CBMs are in the production of oral care products, since CBMs may contribute to disperse the polysaccharides present in the dental plaque; in the baking industry, as CBMs can retard staling and aging of baked bread; in diagnosis test kits (Levy and Shoseyov, 2002) and as part of a system for parenteral vaccination of fish (Maurice *et al.*, 2003). Since CBMs specificity for

carbohydrate ligands is high, these modules have been used as molecular probes for the analysis for plant cell wall polyssacharides (McCartney *et al.*, 2004). Furthermore, Shoseyov *et al.* (2001) also showed that CBMs could modulate plant growth on transgenic plants. The introduction of family 3 CBM gene from *Clostridium cellulovorans* in poplars (*Populus tremula*) made the transgenic plants grow faster than the wild-type, showing an increase in fiber cell length and in the average degree of polymerization of cellulose (Levy and Shoseyov, 2002). Environmental technology has also been exploring the function of CBMs. Several authors fused CBMs with decontamination enzymes and the fusion molecules, attached to cellulose supports successfully degrade toxic compounds (Richins *et al.*, 2000; Wang *et al.*, 2002; Xu *et al.*, 2002).

Although the mechanisms by which these enzymes act have been extensively studied over the last decades, much is still unknown. Whether it is for use in animal nutrition or to develop and/or optimize novel or current biotechnological applications, it is crucial that we improve our knowledge on the processes involved in plant cell wall hydrolysis. This work aims to provide insightfull contributions in this area. This thesis is divided in 6 Chapters. The scientific background (Chapter 1) begins with a general review of plant cell wall structure, with particular focus on cellulose, xyloglucan and galactomannan. Subsequently, attention is focused into the cellulosome, an enzymatic multi-enzyme complex of cellulases and hemicellulases that is usually produced by anaerobic bacteria. *Clostridium thermocellum*'s cellulosome is described and the complexity and functionality of this macromolecule is highlighted. In the following subchapters the role of GHs, CEs and Carbohydrate-Binding Modules (CBMs) in plant cell wall degradation is reviewed. Finally, the chapter finishes with the clear identification of the objectives of this project. In Chapter 2, we address several unresolved questions concerning the mechanism of plant cell wall hydrolysis by microbial multi-enzyme complexes.The crystal structure and biochemical properties of the N-terminal carbohydrate esterase domain of *Clostridium thermocellum Ct*Ces3-1 were determined. In Chapter 3, four carbohydrate-biding modules (CBMs) of family 35 that display specificity for Δ4,5-anhydrogalacturonic acid (Δ4,5-GalA) were characterized, although two of the proteins also interact with glucuronic acid (GlcA). We present the biochemical properties of a CBM6 (CBM6-1) from *Cellvibrio mixtus Cm*Cel5A in Chapter 4 and a novel CBM within the large *C. thermocellum* cellulosomal protein Cthe_2193 (*Ct*CBM55), which is the founder member of a new CBM family in

Chapter 5. Finally Chapter 6, contains a general discussion with the most importante conclusions from each chapter.

CHAPTER 1 Scientific background and objectives

1.1. THE PLANT CELL WALL

Plants present two cell wall types that differ in function and composition: the primary and secondary cell walls. Primary walls surround growing and dividing plant cells, providing mechanical strength but allowing the cell to expand. In contrast, secondary walls are much thicker and stronger and are deposited only when cells have ceased to grow. The secondary walls are strengthened by the incorporation of lignin, a phenolic polymer, which cements and anchors the cellulose microfibrils and other matrix polysaccharides and thus stiffens the walls, preventing biochemical degradation and physical damage (Knudsen, 1997).

The primary wall is a composite polymeric structure in which crystalline cellulose microfibrils are embedded in a complex, highly hydrated, and less ordered polysaccharide matrix (Figure 1.1). The most important components of this matrix are hemicelluloses and pectin. Cellulose is present as long unbranched fibrils composed of approximately 30 to 36 hydrogen-bonded chains of β-1,4-glucose (Somerville *et al.*, 2004). Hemicelluloses are branched polysaccharides containing backbones of neutral sugars, like β-D-hexosyl or β-D-pentosyl residues, that can form hydrogen bonds to the surface of cellulose fibrils. Pectins form a gel phase in which the cellulose-hemicellulose network is embedded and uronic acids are major components of these molecules. The simplest of these is homogalacturonan (HG), an unbranched polymer of α-1,4-D-galacturonic acid. Rhamnogalacturonan I (RGI) has a backbone composed of alternating α-1,2-L-rhamnose-α-1,4-D-galacturonic acid decorated primarily with arabinan and galactan side chains. It has recently been suggested that RGI functions as a scaffold to which other pectins, such as rhamnogalacturonan II (RGII) and HG, are covalently attached as side chains (Somerville *et al.*, 2004).

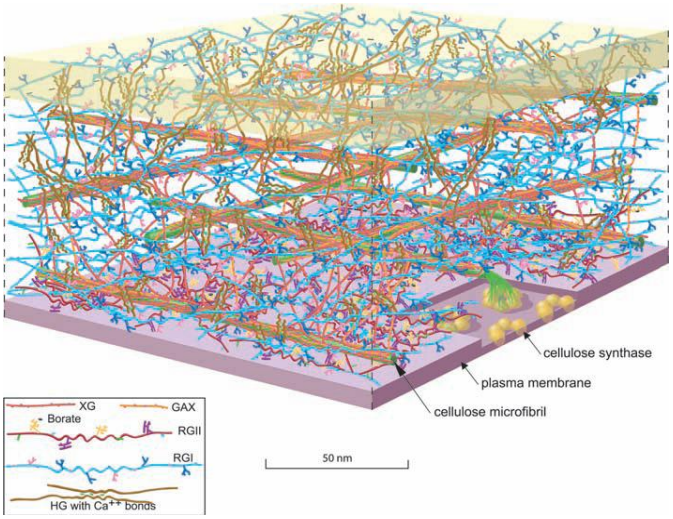


Figure 1.1 Scale model of the *Arabidopsis* leaf primary cell wall.
The amount of the various polymers is shown based approximately on their ratio to the amount of cellulose. The amount of cellulose, relative to a living cell, was reduced for clarity. Because of the exaggerated distance between microfibrils, the hemicellulose cross-links [shown in dark orange (xyloglucan, XG) or light orange (glucuronoarabinoxylan, GAX)] are abnormally extended. Also, recent solidstate NMR studies have suggested that, in some plants, only about 8% of the surface of the cellulose microfibrils is occluded by XG. The figure is an elaboration of a model originally presented by McCann and Roberts. The figure was rendered by Abbey Ryan.

Several models, which reflect the interactions established between its components, have been proposed to explain the organization of plant cell walls. In the seventies, Keegstra *et al.* (1973) proposed that polymers from the matrix (xylan, xyloglucan, pectic polysaccharides and structural proteins) were covalently linked and formed a very large macromolecular network. In this model, cellulose fibres are connected through hydrogen-bonding to xyloglucan chains (Cosgrove, 2001). An alternative model was subsequently proposed by Hayashi (1989) and Fry (1989) that suggested that single xyloglucan chains fill the gap between cellulose microfibrils and tether them together; the pectin polysaccharides and structural proteins occupy the space between xyloglucan chains. Although this is presently the most popular model, two other models have, more recently, been proposed: the multicoat model described by Talbott and Ray (1992), in which each cellulose microfibril is coated with successively looser layers of matrix polysaccharides; and the stratified wall model described by Ha *et al.* (1997), in which cellulose-xyloglucan layers are separated by pectic polysaccharides. All these models have in common the concept that cellulose fibrils are coated with xyloglucan (Cosgrove, 2001). In summary, cellulose microfibrils are linked together by non-covalent interactions with matrix polysaccharides, which determine most of the physical properties of the cell wall. The complexity of the cell-wall network allows for

many potential sites where loosening and expansion might be initiated (Cosgrove, 2005).

1.1.1. Cellulose

The primary structure of cellulose is of an unbranched β1,4-linked β-D-glucan the structural repeat in the polymer being the disaccharide cellobiose due to the 180º rotation of the glucose moieties inside the carbohydrate (Figure 1.2). Many parallel glucans interact due to an extensive hydrogen bond network to form a crystalline microfibril that is mechanically strong and highly resistant to enzymatic attack (an almost ideal scaffold material). These long, crystalline ribbons are 3–5 nm wide and, in growing cells, are aligned with each other, giving a structural bias to the cell wall (Cosgrove, 2005).

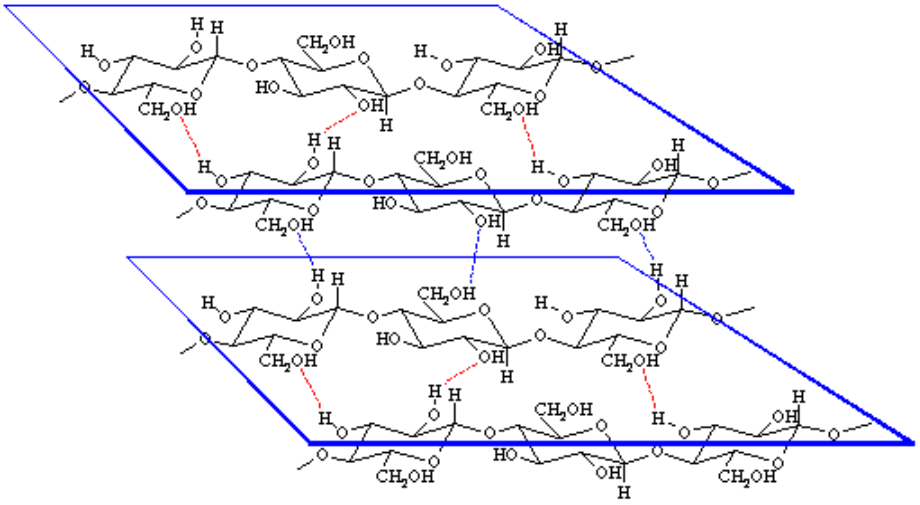


Figure 1.2 Structure of cellulose.
The cellulose is first sandwiched together in sheets. In these sheets, the cellulose forms hydrogen bonds to its adjacent neighbours, increasing mechanical strength and enzymatic recalcitrance. Adapted from http://www.biologie.uni-hamburg.de/b-online/e26/26a.htm.

In nature, cellulose chains have a degree of polymerization of approximately 10,000 glucopyranose units in wood cellulose and 15,000 in native cotton cellulose. There is some evidence for a lower degree of polymerization in primary cell walls as compared with secondary cell walls (O'Sullivan, 1997). Natural crystalline cellulose is named cellulose I and comprises the two forms Iα and Iβ, in which these chains lie parallel (Jamal *et al.*, 2004), although these cellulose architectures are not the most stable form of the polysaccharide (O'Sullivan, 1997). Celluloses produced by primitive

organisms were said to have the Iα component dominant, while those produced by the higher plants have the Iβ form dominant (O'Sullivan, 1997). Many non-natural forms of cellulose and crystalline arrays of cellooligosaccharides form cellulose II in which the chains lie anti-parallel (Jamal *et al.*, 2004). In addition to crystalline regions, many model sources of natural crystalline cellulose appear to contain various proportions of unstructured cellulose, rather loosely termed "amorphous" cellulose. However, amorphous cellulose cannot be considered truly amorphous as, by definition, an amorphous material is one which is formless or lacks a definite shape and areas of amorphous cellulose probably still possess a degree of order (O'Sullivan, 1997).

1.1.2. Xyloglucan

Xyloglucan is the quantitatively predominant hemicellulosic polysaccharide in the primary walls of dicots. It has a cellulosic semirigid backbone of β-1,4-glucan that is substituted with α-1,6-D-xylosyl units which, in a species-dependent manner, is further derivatized with α-L-arabinose or β-D-galactose. Usually, up to 75% of the residues are substituted at O6 with mono-, di-, or triglycoside side chains. As displayed in Figure 1.3, a single letter nomenclature is used to simplify the naming of xyloglucan side chain structures. For example, a capital **G** represents an unbranched glucose residue. A capital **F** represents a glucose residue with a fucose-containing trisaccharide.

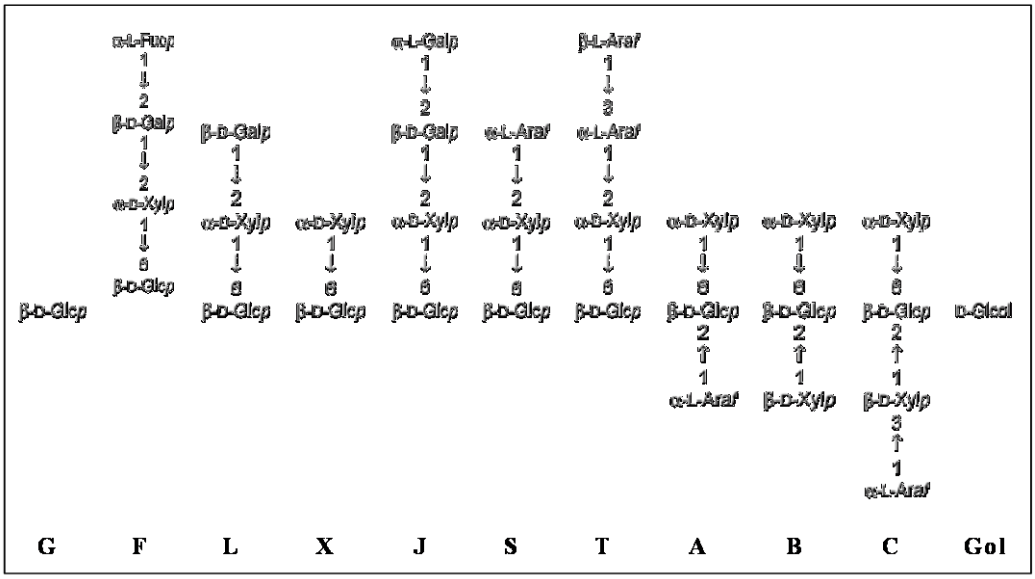


Figure 1.3 Xyloglucan architecture and the one letter code to define polysaccharide structure. Adapted from Fry *et al.* (1993)

Xyloglucans are classified as XXXG-type or XXGG-type based on type of decorations. XXXG have three consecutive backbone residues that are subsituted with xylose and a fourth unbranched backbone sugar (

Figure 1.4). In contrast, XXGG xyloglucans have two consecutive branched backbone residues that are followed by two unbranched backbone residues. It is well established that xyloglucan structures vary between plant species. However, primary walls of several dicotyledons, non-graminaceous monocotyledons and gymnosperms contain fucosylated xyloglucan with a XXFG-type structure.

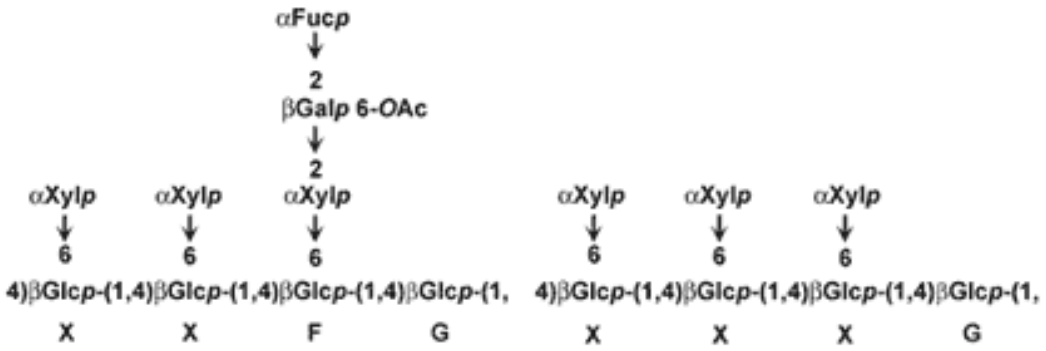


Figure 1.4 Schematic representation of XXXG-type xyloglucans.
Adapted from http://www.ccrc.uga.edu/~mao/xyloglc/Xtext.htm (Mars 2009).

1.1.3. Mannan, glucomannan and galactomannan

The mannose-containing polysaccharides, mannan and glucomannan, are important components of plant biomass. Mannan consists of a backbone of β-1,4-linked mannose residues, whereas glucomannan comprises a heterogeneous polymer of β-1,4-linked glucose and mannose sugars randomly distributed (Brett *et al.*, 1996). The backbone of both mannan and glucomannan can be decorated with α-1,6-linked galactosyl residues, and thus these polysaccharides are often referred to as galactomannan and galactoglucomannan, respectively (Brett *et al.*, 1996). Glucomannan plays a key structural role in the plant cell walls of Angiosperms, while galactomannan is generally found in the cell walls of seeds such as carob, where it functions as a storage polysaccharide (Hogg *et al.*, 2003). Galactoglucomannan is present mainly in softwoods, whereas in hardwoods glucomannan is the most common form (Vries *et al.*, 2001).

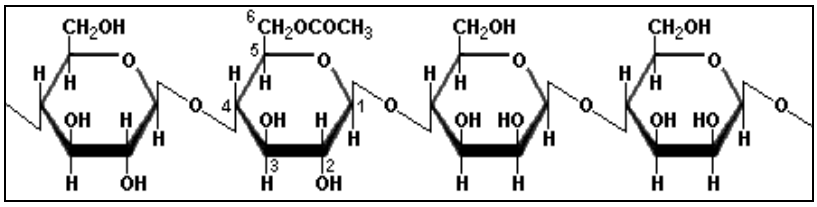


Figure 1.5 A portion (Glucose, Glucose, Manose, Manose) of the glucomannan repeating unit. The second glucose has an acetate group.

Galactomannans are present in several vegetable gums that are used to increase the viscosity of food products. The mannose:galactose ratio varies between different gums like Fenugreek gum (1:1), Guar gum (2:1), Tara gum (3:1), Locust bean gum or Carob gum (4:1). Guar is a legume that has been traditionally cultivated as livestock feed. Approximately 85% of guar gum is guaran, a water soluble polysaccharide consisting of linear chains of mannose with β-1,4 linkages to which galactose units are attached with α-1,6 linkages. The ratio of mannose to galactose is 2:1. Guar gum has five to eight times the thickening power of starch and has many uses in the pharmaceutical industry, as a food stabilizer and as a source of dietary fiber.

1.1.4. Pectins

Pectic substances constitute the major matrix polysaccharides in the middle lamella and primary cell walls of dicotyledonous plants, with a backbone composed of alternating homogalacturonan (smooth regions) and rhamnogalacturonan (hairy regions) (Kofod *et al.*, 1994). Rhamnogalacturan I consists of alternating residues of galacturonic acid and rhamnose, and probably has side branches that contain other pectin domains (Cosgrove, 2005). The latter can have acetylations at the C-2 and C-3 positions, and the removal of such acetyl groups facilitates the action of lyases and hydrolases, since the acetylation sterically hinders the cleavage of the glycoside linkages (Navarro-Fernández *et al.*, 2008). Homogalacturonan comprises a linear chain of galacturonic acid residues, whereas xylogalacturonan is modified by the addition of xylose branches. The carboxyl groups of homogalacturonan and xylogalacturonan are often methyl esterified, a modification that 'blocks' the acidic group and reduces their ability to form gels (Cosgrove, 2005). Rhamnogalacturonan II is a complex pectin domain that contains 11 different sugar residues and forms dimers through borate (B) esters. The neutral arabinans and arabinogalactans are also linked to the acidic pectins and it has been proposed that they promote wall flexibility while they bind to the surface of cellulose.

1.2. HYDROLYSIS OF PLANT CELL WALL POLYSACCHARIDES BY HIGH MOLECULAR MASS MULTI ENZYME COMPLEXES

In nature, different mechanisms have evolved for the hydrolysis of plant cell wall structural carbohydrates. Aerobic bacteria and fungi usually secrete a large repertoire of cellulases and hemicellulases to the extracellular space, since structural carbohydrates cannot enter the microbial cells, which individually bind to each specific target polysaccharide through the action of non-catalytic CBMs. In contrast, several anaerobic bacteria and fungi organize cellulases and hemicellulases in high molecular mass multi enzyme complexes that present one of the finest examples of a naturally evolved nono-machine. These multi enzyme complexes were termed as Cellulosomes due to their very high efficiency for cellulose hydrolysis. Anaerobic bacteria producing cellulosomes were identified in the genera *Clostridium*, *Acetivibrio*, *Bacteroides* and *Ruminococcus* that colonize various environmental niches, including the soil, wood chip piles, sewage, and the rumen. Cellulosomes may be the largest extracellular multienzyme complexes found in nature, since polycellulosomes have been reported to be as large as 100 MDa, although the individual cellulosomes range in mass from about 650,000 Da to 2.5 MDa (Doi *et al.*, 2003). Generally, enzymatic complexes from anaerobic microorganisms are much more elaborated and complex when compared with their aerobic counterparts where enzymes act individually during cell wall hydrolysis. It is possible that the anaerobic environment presents a greater selective pressure for the evolution of a highly efficient machinery (Bayer *et al.*, 2004). Aerobic cellulolytic fungi, for example, are able to produce and secrete copious amounts of free cellulases and hemicellulases which act individually during plant cell wall hydrolysis. When the first attempt to study bacterial anaerobic cellulases and hemicellulases were made, scientists found it rather odd that those enzymes could not be found in the free state (Bayer *et al.*, 2004). Only later it was recognized that this phenomenon resulted from the association of the hydrolytic enzymes in cellulosomes that are usually attached to the bacterial surface. Studies indicating that *C. thermocellum* enzymes were probably associated in a multi-enzyme complex date back from the eighties of the last century (Garcia-Martinez *et al.*, 1980; Lamed and Zeikus, 1980). In these initial studies, Bayer *et al.* (1985; 1986) presented electron microscopy images displaying the *C. thermocellum* cellulosome, suggesting that this multi-enzyme complex appears to be centralized on protuberant structures primarily located on the bacterial surface (Figure 1.6).

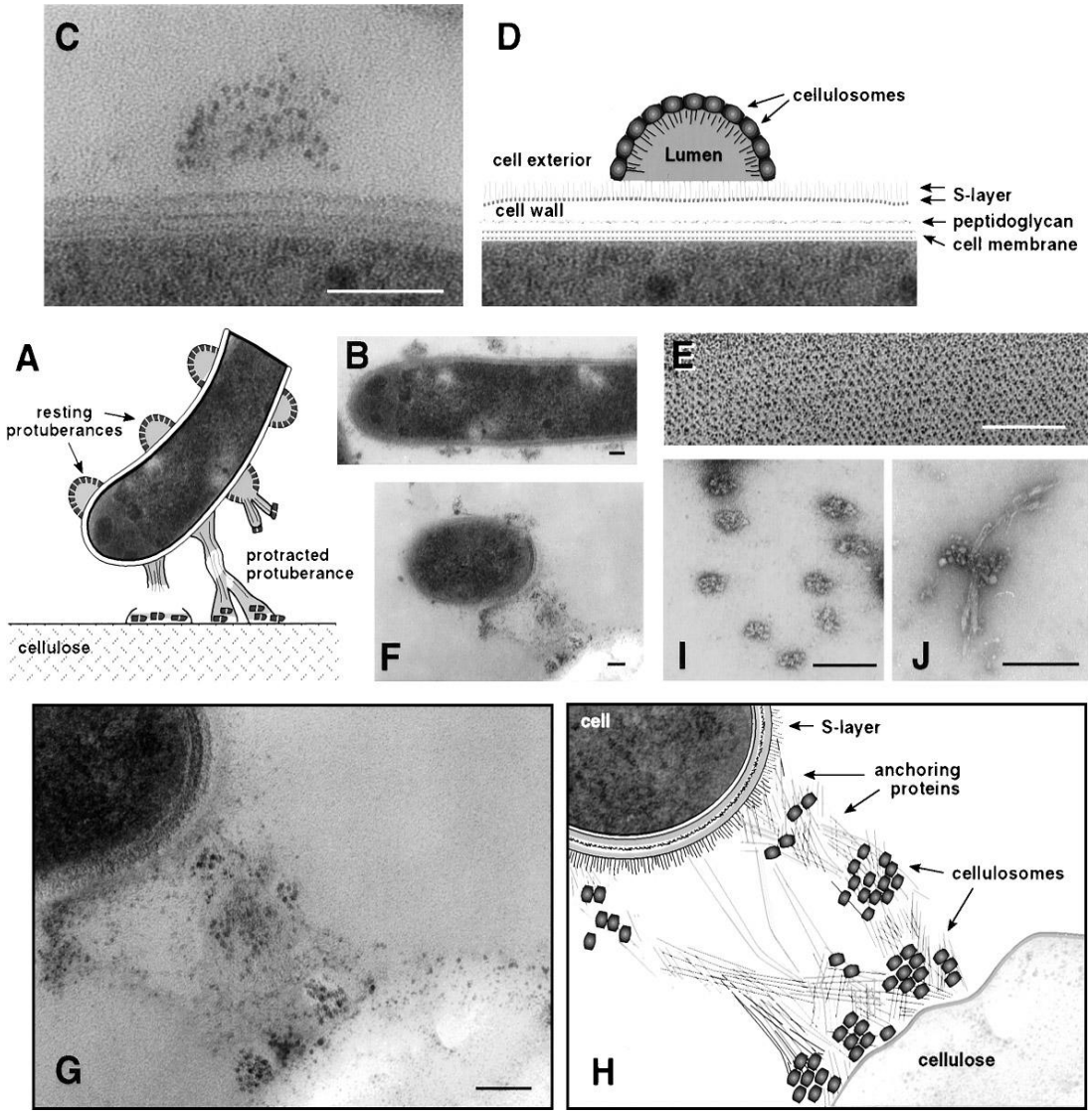


Figure 1.6 Ultrastructure of *Clostridium thermocellum* cell surface and cellulosomes.
(A) Diagrammatic representation of a typical *C. thermocelum* cell bound to cellulose. The cell is intermittently covered with polycellulosomal protuberance-like organelles, some of which are in the resting state while others have protracted upon binding to the substrate. (B) Transmission electron micrograph of a cell in the free state, prior to contact with cellulose. The cell was stained with cellulosome-specific antibody. (C) A high-resolution magnification of a quiescent cellulosome-specific antibody-labeled polycellulosomal protuberance. Note the label on the outer surface of the protuberances. (D) Schematic interpretation of the cell surface shown in C. (E) A rotary-shadowed, transmission electron micrograph of a cell envelope fragment of *C. thermocellum*. (F) Transmission electron micrograph of a cellulose-bound cell, stained with cellulosome-specific antibody. Upon binding to the substrate, the polycellulosomal organelle has unfurled. (G) A high-resolution magnification of a protracted, antibody-labeled polycellulosomal protuberance. The cellulosome-specific label is mainly associated with the cellulose surface and connected to the cell via extended fibrous material. Compare with micrograph in C. (H) Schematic interpretation of the cellulose-bound cell surface shown in G. (I) Transmission electron micrograph of negatively stained, purified cellulosomes in the absence of cellulose. Note multicomponent nature of the cellulosomes. (J) Similar micrograph of cellulosome bound to fibers of bacterial microcrystalline cellulose. Bars in B, C, E–G, 100 nm; in I and J, 50 nm. Adapted from Bayer *et al.* (1998).

The initial definition of the cellulosome concept was based on studies of the cellulase system produced by the anaerobic thermophilic cellulolytic bacterium *Clostridium thermocellum*, where adherence of the microorganism to the cellulose was proven (Bayer *et al.*, 1983). It is now recognized that cellulosomes may actively degrade other plant cell wall components and not only cellulose but also a variety of hemicellulases, by incorporating carbohydrate esterases and pectate lyases in this multi-enzyme complex (Bayer *et al.*, 2004).

Typical bacterial cellulosomes are composed of a scaffolding protein, termed CipA in *C. thermocellum*, that contains a powerful cellulose-binding module (CBM), usually of family 3 (see below), and a number of cohesin modules which tightly bind to complementary dockerin modules borne by the catalytic cellulosomal subunits (Bayer *et al.*, 2004). Moreover, the scaffoldin contains a C-terminal modular dyad that includes a dockerin that displays a divergent type of specificity to that of the enzyme-borne dockerins. Instead, CipA dockerin, binds selectively to cohesins located on a set of *anchoring scaffoldins*, each of which contains a C-terminal S-layer homology module that mediates attachment to the bacterial cell surface (Bayer *et al.*, 2008). Due to the different specificities of the cohesin–dockerin pairings, the cohesins of the primary scaffoldin were termed type-I cohesins, whereas those of the anchoring scaffoldins were designated type-II, while dockerin type is set according to that of the cohesin with which it interacts (Bayer *et al.*, 2008). The family-3 CBM (CBM3) located in scaffoldins binds strongly to crystalline cellulose, which accounts for the primary targeting of the cellulosome to its substrate (Tomme *et al.*, 2003). The three-dimensional crystal structure of *C. thermocellum* CipA CBM3 has been determined, and a general molecular mechanism for binding to the cellulosic surface has been proposed, which is described below (Bayer *et al.*, 2008). A schematic representation of *Clostridium thermocellum* cellulosome is shown in Figure 1.7. Other bacteria, like *Acetivibrio cellulolyticus*, *Bacteroides cellulosolvens* and *Ruminococcus flavefaciens* also seem to have dockerin-containing scaffoldins and SLH-bearing anchoring proteins (Ding *et al.* 1999; Ding *et al.*, 2001; Xu *et al.*, 2004).

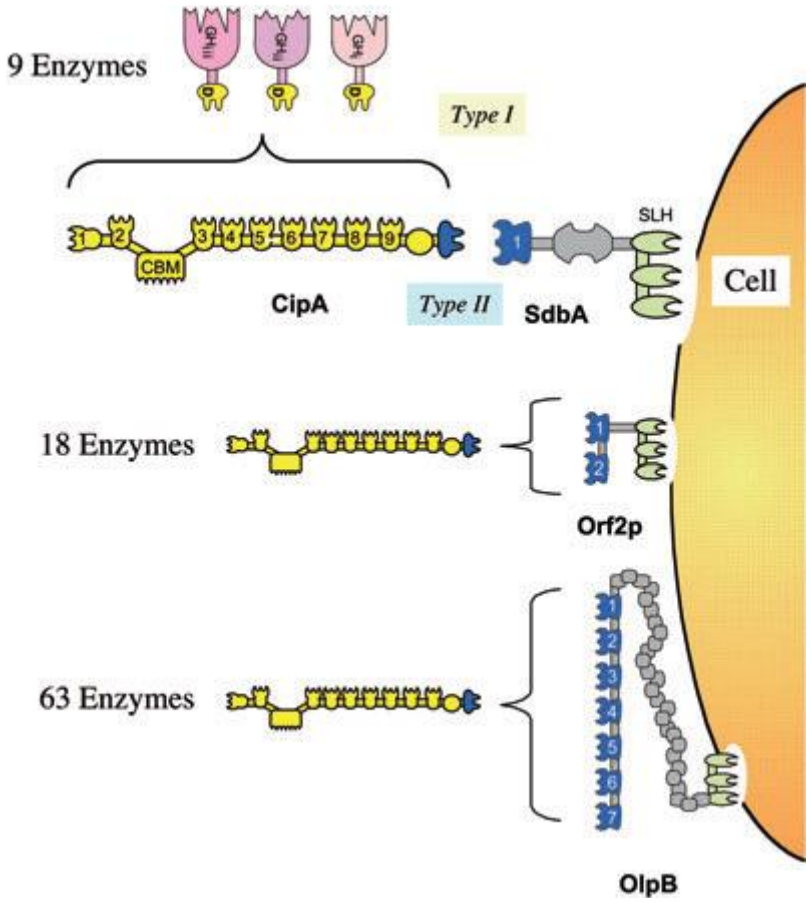


Figure 1.7 Schematic representation of *Clostridium thermocellum* cellulosome and mechanisms of cellulosome cell surface attachment.
Schematic representation of the supramolecular Lego-like architecture of the *Clostridium thermocellum* cellulosome paradigm and its disposition on the bacterial cell surface. Dockerin-containing enzymes bind selectively to any of the nine type-I cohesins (enumerated) of the primary CipA scaffoldin. In turn, the terminal X–dockerin dyad of the CipA scaffoldin binds to the type-II cohesins of anchoring scaffoldins SdbA, Orf2p, or OlpB, each of which is connected to the cell surface via an S-layer homology module (SLH). The cellulose-binding module of the primary scaffoldin binds the cellulosome complex and the attached cell to the cellulosic substrate. Adaped from Bayer *et al.* (2008).

Thus, the scaffoldin subunit is responsible for three critical functions of the *C. thermocellum* cellulosome: (i) it integrates nine different dockerin-bearing enzyme subunits into the cellulosome complex by virtue of its nine cohesin modules, (ii) it participates in the attachment of the cellulosome to the bacterial cell surface by virtue of its C-terminal dockerin module that binds selectively to cohesins of the anchoring scaffoldins, and (iii) it serves to target the cellulosomal enzymes as well as the entire cell to the cellulosic substrate by virtue of its CBM3 (Bayer *et al.*, 2008).

Cellulosomal enzymes contain minimally a catalytic domain and a 70-amino-acid-long duplicated sequence that binds scaffoldin cohesin domains and which was termed dockerin. The presence of a dockerin in an enzyme usually indicates that it is a

cellulosomal enzyme, since the dockerin interacts with the cohesins of the scaffolding protein to form the enzyme complex. Interestingly, most of the known *C. thermocellum* cellulosomal cellulases and hemicellulases contain additional CBMs, which make the cellulosomal enzymes similar to their free enzyme analogs. Rather than primary binding of the crystalline cellulosic substrate, the latter modules are considered to perform functions of a subtler nature. In some cases, such as the hemicellulases, an enzyme-borne CBM could presumably direct the cellulosome-bound enzyme to the favored site in the hemicellulosic polysaccharide substrate. Finally, in some instances, cellulosomal enzymes from *C. thermocellum* contain more than one catalytic domain, perhaps indicating a close cooperation between two particular types of catalytic domain, which may promote synergistic action at a given site on the cellulosic or hemicellulosic substrate (Bayer *et al.,* 2004).

There has been a recent interest in constructing designer mini-cellulosomes for several purposes, including studying the functions of cohesins, analyzing synergy between various cellulosomal enzymes and improving the efficiency of cellulosomes (Doi *et al.*, 2003). There is also an increasing interest in developing *in vivo* systems in various bacteria in order to use relatively inexpensive biomass or agricultural wastes as a substrate to obtain valuable products (Doi *et al.*, 2003). One example of this type of research is to insert cellulosomal genes into organisms that already are able to produce valuable products, such as ethanol, butanol, or amino acids. Cellulosomes are important in that they can degrade crystalline forms of cellulose, which are recalcitrant to many enzymes that can degrade soluble or amorphous forms of cellulose. Furthermore, cellulosomes can degrade hemicelluloses and pectin, two other major components of plant cell wall materials.

1.2.1. Structure and funtion of Carbohydrate-Active Enzymes

In general, enzymes that degrade plant cell wall polysaccharides display a modular architecture, containing one more catalytic domains bound, through flexible linker sequences, to one or more non- catalytic modules. Most of the characterized non-catalytic binding modules directing the appended catalytic regions to their target substrates and are known as Carbohydrate-Binding Modules (CBMs). Below, a detailed description of catalytic and non-catalytic modules found in cellulases and hemicellulases is presented.

1.2.1.1. Glycoside hydrolases

The extensive variety of carbohydrates found in plant cell walls requires for their degradation a large multiplicity of enzymes that are thus involved in their metabolism. Glycoside hydrolases (EC 3.2.1.x) are key enzymes in carbohydrate hydrolysis and are found in the three major kingdoms (archaebacteria, eubacteria and eukaryotes) (Henrissat *et al.*, 1991). Glycoside hydrolases cleave the glycosidic bond between two carbohydrates or between a carbohydrate and a non-carbohydrate moiety. Many of these are enzymes produced by saprophytic microorganisms for the degradation of structural polysaccharides (Henrissat *et al.*, 1998).

1.2.1.1.1. Nomenclature of glycoside hydrolases

The International Union of Biochemistry and Molecular Biology enzyme nomenclature (IUB-MB; 1984) is based on the type of reaction that enzymes catalyse and on their substrate-specificity. For GHs (EC 3.2.1.x), the first three digits indicate enzymes hydrolysing *O*-Glycoside linkages, whereas the last number indicates the substrate and sometimes reflects the molecular mechanism (Henrissat *et al.*, 1991). This classification avoids ambiguities and the proliferation of trivial names, been for that reason very useful. However, in the case of glycoside hydrolases, such classification does not necessarily reflect the structural features of the enzymes and it is also not appropriate for enzymes showing broad substrate specificity (Henrissat *et al.*, 1991). In 1991, Henrissat and colleagues proposed a new classification system for glycoside hydrolases, where enzymes were organized in families based on primary sequence similarities. Since there is a direct relationship between the amino acid sequence and the folding of an enzyme, such classification is expected to: (i) reflect the structural features of these enzymes better than substrate specificity alone; (ii) help to reveal the evolutionary relationships between these enzymes; and (iii) provide a convenient tool to derive mechanistic information from the protein sequence data.

The Henrissat classification does not substitute but rather complements the IUB-MB nomenclature. An advantage of this classification is that a protein, a translated DNA sequence, or even a domain can be classified before even knowing its biochemical properties. In addition, it overcomes the problems associated with the classification of enzymes consisting of several catalytic domains. Significant sequence similarity is a strong indication of folding similarity and, therefore, it was found that members of one family most likely share the same folding characteristics, enabling homology modelling

if the three-dimensional structure of one member is known (Henrissat *et al.*, 1991). According to the Henrissat classification, the GHs catalytic modules are currently classified into 114 different families based on amino acid sequence similarities (May 2009). Because the fold of proteins is better conserved than their sequences, some of the families can be grouped in 14 'clans'.

Table 1.1 Glycoside hydrolases fold superfamilies.
Adapted from http://www.cazy.org

Clans of Related Families	Protein fold	Glycoside Hydrolase Families
GH-A	$(\beta/\alpha)_8$	1 2 5 10 17 26 30 35 39 42 50 51 53 59 72 79 86 113
GH-B	β-jelly roll	7 16
GH-C	β-jelly roll	11 12
GH-D	$(\beta/\alpha)_8$	27 31 36
GH-E	6-fold β-propeller	33 34 83
GH-F	5-fold β-propeller	43 62
GH-G	$(\alpha/\alpha)_6$	37 63
GH-H	$(\beta/\alpha)_8$	13 70 77
GH-I	α+β	24 46 80
GH-J	5-fold β-propeller	32 68
GH-K	$(\beta/\alpha)_8$	18 20 85
GH-L	$(\alpha/\alpha)_6$	15 65
GH-M	$(\alpha/\alpha)_6$	8 48
GH-N	β-helix	28 49

Henrissat and colleagues (1998) also proposed, that enzymes could be named according to their target substrate, following the three-letter standard used in bacterial genetics for genes, although this designation would comprehend the family to which the enzyme belongs. Thus, a family 5 GH will be named Cel5 or Man5, depending on its substrate, which could be cellulose or mannan, respectively. If an organism produces multiple enzymes from the same family displaying similar substrate specificities, then they would be designated Cel5A and Cel5B, and so on, where the letters after the family number correspond to the order in which the enzymes were first reported (Henrissat *et al.*, 1998). If an enzyme contains more than two catalytic domains, the designation would include all of them (Henrissat *et al.*, 1998). For example, endoglucanase CelA from *Caldocellulosiruptor saccharolyticum* is composed of two GHs, one from family 9 and the other from family 48. So, according to this new nomenclature, the enzyme is named *Cs*Cel9A-Cel48A, written in the conventional sense from the amino to the carboxyl terminus. The microorganism abbreviation is also included, before the enzyme name, to differentiate similar enzymes of different origins.

1.2.1.2. Carbohydrate Esterases

As with the classification of glycoside hydrolases, the Carbohydrate-Active Enzyme Server (www.cazy.org; July 2009) provides a classification of carbohydrate esterases based upon amino acid sequence similarities. Currently, 16 such families have been described (May 2009), ten of which contain enzymes involved in the degradation of plant-cell wall polysaccharides such as xylan, pectin and rhamnogalacturonan. Biologically, these enzymes are involved in the removal of O-(ester) and N-(acetyl) moieties from carbohydrates and indeed, sugar deacetylases display similar catalytic strategies to those employed by more classical esterases and peptidases. In fact, the similarity of different sugar and non-sugar esterases, and the promiscuity of their substrate specificity make the carbohydrate esterase classification in many cases less insightful and predictive than the hydrolase, transferase and lyase families (Davies *et al*, 2005).

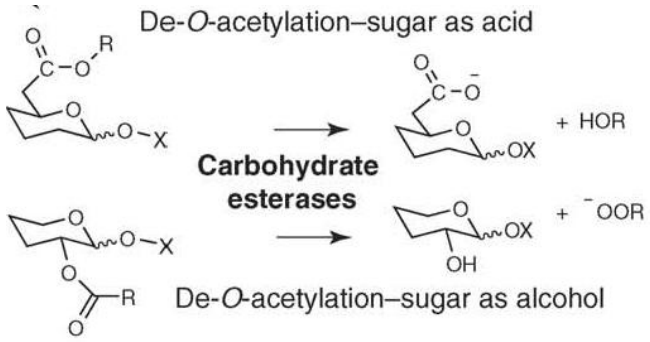


Figure 1.8 mechanism of action of carbohydrate esterase.
Carbohydrate esterases perform the O or N deacetylation of acetylated sugars.

The majority of carbohydrate esterases/deacetylases whose structures have been reported display a classical β/α/β 'serine protease' fold, as revealed by three-dimensional structures of enzymes from families CE-1 (bacterial ferulate esterases, Prates *et al.*, 2001), CE-5 (acetyl xylan esterases, Molgaard *et al.*, 2002), CE-7 (multifunctional and xylooligosaccharide deacetylases, Vicent *et al*, 2003), the plethora of enzymes from family CE-10 (www.cazy.org) and the non-classified fungal ferulate esterases (Hermoso *et al.*, 2004). There is also a small deviation from this canonical fold in the CE-12 rhamnogalacturonan acetylesterase (Molgard *et al.*, 2004). In contrast, the pectin methylesterases from family CE-8 use a different mechanism, in which a twin-aspartate catalytic centre is grafted onto a right-handed parallel β-helix and these enzymes may also be considered slightly unusual in that it is the sugar that forms the

acid, rather than the 'R' group (see Figure 1.8) (Jenkins *et al.*,2001). Some sugar deacetylase structures have also revealed both single and double metal ion catalytic centres. For example, the LpxC zinc-dependent UDP-3-O-acetyl-Nacetylglucosamine deacetylases from family CE-11, which presents a classical zinc hydrolase site on a novel α/β framework (Coggins *et al.*, 2003; Whittington *et al.*, 2003); the single-zinc CE-14 N-acetyl-1-D-myo-inosityl-2-amino-2-deoxy-a-D-glucopyranoside deacetylase (Maynes *et al.*, 2003); the family CE-4 deacetylases (Blair *et al.*, 2005), sometimes refered to as 'NodB homologs', whose members are involved in the deacetylation, amongst other things, of peptidoglycan, chitin, rhizobial Nod factors and xylan; and the twin-metal, urease-like CE-9 N-acetylglucosamine-6-phosphate deacetylase (Vincent *et al.*, 2004).

1.2.1.3. Carbohydrate-Binding Modules

Enzymes that digest plant cell wall polysaccharides generally contain non-catalytic domains, CBMs, that function by anchoring the enzyme to the substrate, thus potentiating catalytic activity (Tunnicliffe et al., 2005). CBMs were initially classified as cellulose binding domains (CBDs), based on the initial discovery of several modules that bind cellulose (Shoseyov et al., 2006). However, more and more modules in carbohydrate-active enzymes that bind carbohydrates other than cellulose are being found and these domains have assumed the more general terminology of Carbohydrate-Binding Modules (CBMs).

CBMs have three general roles with respect to the function of their cognate catalytic modules: (i) a proximity effect, (ii) a targeting function and (iii) a disruptive function (Boraston *et al.*, 2004). Through their sugar-binding activity, CBMs concentrate enzymes on to the polysaccharide substrates. It is thought that maintaining the enzyme in proximity of the insoluble substrate (i.e. increasing the concentration of the enzyme on the surface of the substrate) leads to a more rapid degradation of the polysaccharide (Bolam *et al.*, 1998), therefore resulting in improved enzyme efficiency. There are numerous examples in the literature where proteolytic excision or genetic truncation of CBMs from the catalytic modules results in significant decreases in the activity of the enzymes on insoluble, but not soluble polysaccharides (Tomme *et al.*, 1988; Gilkes *et al.*, 1988; Bolam *et al.*, 1998; Charnock *et al.*, 2000; Ali *et al.*, 2001; Zverlov *et al.*, 2001; Boraston *et al.*, 2003; Hall *et al.*, 1995). It should be pointed out

that there are CBMs that have become components of the substrate-binding sites of glycoside hydrolases, and are pivotal to the substrate specificity and mode of action of the enzymes. Therefore, a CBM22 was shown to change the specificity of a glycoside hydrolase family 10 xylanase, such that it displayed primarily *β*-1,4-*β*-1,3-glucanase activity (Araki *et al*., 2004; Boraston *et al*., 2004).

CBMs display substantial variation in ligand specificity. There are CBMs that interact with crystalline cellulose, non-crystalline cellulose, chitin, *β*-1,3-glucans and *β*-1,3-1,4-mixed linkage glucans, xylan, mannan, galactan and starch, while some CBMs display 'lectin-like' specificity and bind to a variety of cell-surface glycans. In general, CBMs are appended (by relatively unstructured linker sequences) to catalytic domains that degrade structural polysaccharides. Although many of these modules target components of the plant cell wall, several CBM families contain proteins that bind to insoluble storage polysaccharides such as starch and glycogen.

CBMs have also been found not only in cellulases and xylanases but also in several hemicellulases displaying a variety of substrate specificities. For example, in *Trichoderma reesei*, CBMs were identified in endomannanase and acetylxylanesterases (Stålbrand *et al.*, 1995; Margolles-Clark *et al.*, 1996; Shoseyov *et al.*, 2006). Other examples of enzymes containing CBMs in their structure are an esterase from *penicillium funiculosum* (Kroon *et al.*, 2000), an isomaltodextranase from *arthrobacter globiformis* (Hatada *et al.*, 2004), arabinofuranosidases from *Aspergillus kawachii* and from *Cellvibrio japonicas* (Bolam *et al.*, 2001; Miyanaga *et al.*, 2004), a pectate lyase from *Pseudomonas cellulose* (Brown *et al.*, 2001), a β-glucosidas from *Phanerochaete chrysosporium* (Lymar *et al.*, 1995) and a dextranase from *Paenibacillus sp* (Finnegan *et al.*, 2005).

Expansins and lectins are homologous to CBMs in the sense that they also bind carbohydrates. Expansins seem to play a role in nonhydrolytic cell wall expansion and possess cellulose binding capabilities *in vitro* (Shoseyov *et al.*, 2006). Lectins are involved in the recognition of glycoproteins at cell surfaces (McGreal *et al.*, 2004). Therefore, like CBMs, lectins are proteins that recognize carbohydrates and fulfil an importante role in the immuse system that has been extensively studied over the last years.

CBMs are also part of the scaffoldin subunits that organize the catalytic subunits in the cellulosome, like CBM3 from *Clostridium thermocellum* scaffoldin CipA. It is clearly established that removal of CBMs from enzymes, or from scaffoldins,

dramatically reduces the enzymatic activity of the associated catalytic modules (Tomme *et al.*, 1988; Gilkes *et al.*, 1988; Hall *et al.*, 1995; Bolam *et al.*, 1998; Zverlov *et al.*, 2001; Boraston *et al.*, 2003; Ali *et al.*, 2005).

1.2.1.3.1. Nomenclature and types of CBMs

A CBM is defined as a contiguous amino acid sequence within a protein with a discrete fold having carbohydrate binding activity (Shoseyov *et al.*, 2006). To date, more than 300 putative sequences in more than 50 different species have been identified, and like GHs, they are divided into families according to the homology shared at the level of their primary amino acid sequences. So far, CBMs have been grouped into 53 families (http://www.cazy.org/fam/acc_CBM.html; January 2009) in a systematic nomenclature similar to the one adopted for GHs. At its simplest, a CBM is named by its family, e.g. the family 6 CBM from *Clostridium thermocellum* XynZ would be called CBM6, but one may also include the organism and even the enzyme from which it is derived to improve clarity. Thus this CBM6 may be defined as *Ct*CBM6 or *Ct*Xyn*Z*CBM6. If glycoside hydrolases contain tandem CBMs belonging to the same family, a number corresponding to the position of the CBM in the enzyme relative to the N-terminus is included (Boraston *et al.*, 2004). For example, *Clostridium thermocellum* enzyme Cthe_2137 contains two CBMs from family 35 and thus first CBM is referred to as *Ct*CBM35-1 and the second as *Ct*CBM35-2.

A classification of CBM families by their structural fold, similarly to what was described for the GHs superfamily classification, was proposed by Boraston *et al.* (2004). Thus CBM families were classified into seven structural family folds (β-sandwich, β-trefoil, cystein knot, unique, OB fold, hevein fold and hevein-like fold). However, the dominant fold among CBMs is the β-sandwich fold, family fold 1 (Table 1.2), and it comprises two *β*-sheets, each consisting of three to six antiparallel *β*-strands (Boraston *et al.*, 2004). CBMs share this fold with plant legume lectins and animal galectins, pentraxins, spermadhesins, calnexin, although no CBM has significant amino acid sequence similarity with these other proteins (Boraston *et al.*, 2004). With the exception of *Cf*CBM2a from *Cellulomonas fimi* xylanase 10A, all of the *β*-sandwich CBMs have at least one bound metal atom. Generally, these metal ions appear to have a structural role; however in *Pp*CBM36 from *Paenibacillus polymyxa* Xyn43A (Figure 1.9) the ligand interaction is mediated by a calcium atom (Jamal-Talabani *et al.*, 2004). In the beginning of 2004, the majority of CBM structures with a β-sandwich motif had a

β-jelly roll fold. However, in the last few years, a couple of CBMs with β-sandwich structures have been shown to have an immunoglobulin fold (Hashimoto *et al.*, 2006). Thus, the β-sandwich CBM family is divided into two fold sub-families: β-jelly roll and immunoglobulin.

Table 1.2 Classification of CBM fold families.
Adapted from Boraston *et al.* (2004) and Hashimoto (2006).

Fold family	Fold		CBM families
1	β-Sandwich	β-jelly immunoglobulin	2,3,4,6,11,15,17,22,27,28,29,30,32,35,36,44 9,20,25,26,31,33,34
2	β-Trefoil		13, 42
3	Cysteine knot		1
4	Unique		5,12
5	OB fold		10
6	Hevein fold		18
7	Unique; conteins hevein-like fold		14

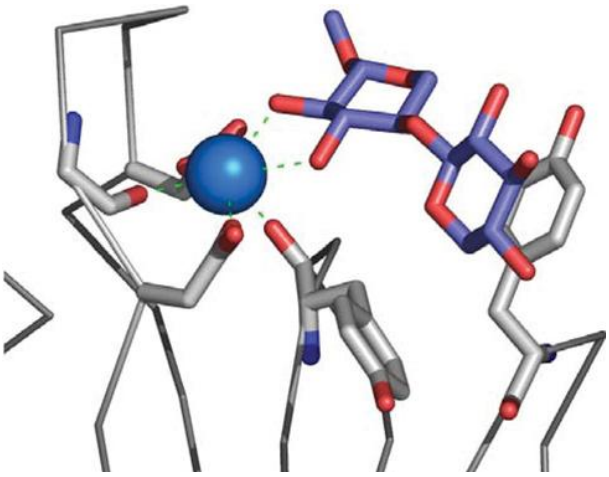


Figure 1.9 The role of calcium in Xylan recognition by the family 36 CBM from *Paenibacillus polymyxa* Xyn43A.
The calcium atom bound in the binding site is shown as a blue sphere; the amino acid residues involved in co-ordinating the calcium atom and binding the sugar are presented in grey 'liquorice' representation; the bound xylo-oligosaccharide is shown in blue 'liquorice' representation and the C-α backbone is represented as grey cylinders. Adapted from Boraston *et al.* (2004).

Fold family 2, which consists of a β-trefoil, is commonly associated with the ricin toxin β-chain. This fold contains 12 strands of β-sheet, forming six hairpin turns. A β-barrel structure is formed by six of the strands, attendant with three hairpin turns and the other three hairpin turns form a triangular cap on one end of the β-barrel called the 'hairpin triplet' (Boraston *et al.*, 2004). The subunit of this fold, called here a trefoil

domain, is a contiguous amino acid sequence with four *β*-strand, two-hairpin structure having a trefoil shape; each trefoil domain contributes one hairpin (two *β*-strands) to the *β*-barrel and one hairpin to the hairpin triplet (Boraston *et al*., 2004). Family 13 CBMs was initially the only group to show this fold, however subsequent determination of the crystal structure of family 42 CBMs revealed that these CBMs also adopt the β-trefoil fold (Hashimoto *et al.*, 2006).

Members of fold families 3, 4 and 5, which are small 30–60 amino-acid polypeptides containing only *β*-sheet and coil, show less diversity in their ligand specificities and appear to be specialized for the recognition of cellulose and/or chitin (Boraston *et al*., 2004). The majority of these CBMs have planar carbohydrate-binding sites comprising aromatic residues (Boraston *et al*., 2004). Hevein domains, fold super-families 6 and 7 (Table 1.2), are small CBMs with approimately. 40 amino acids, and originally they were identified in plants as chitin-binding proteins. The fold comprises predominantly a coil, but does have two small *β*-sheets and a small region of helix (Boraston *et al*., 2004). The minimal hevein fold is found in family 18 CBMs which is classified as CBM fold family 6 (Table 1.2). Family 14 CBMs incorporate aspects of the hevein domain, however, this family also have a fusion of this fold with a small β-sheet structure, which leads to a new classification as a separate fold family, family fold 7 (Table 1.2) (Boraston *et al*., 2004).

Although CBM families can be grouped into fold super-families based on protein fold conservation, such groupings do not predict CBM function. Therefore, a classification based on ligand binding properties displayed by CBMs has also been proposed, where CBMs are grouped into three types: ‘surface-binding’ CBMs (type A), ‘glycan chain-binding’ CBMs (type B) and ‘small sugar-binding’ CBMs (type C) (Hashimoto *et al.*, 2006), as depicted in Table 1.3.

Table 1.3 The different types of CBMs.
The table was adapted from Boraston *et al.* (2004) and http://www.cazy.org.

CBM Type	Fold family	CBM families
A	1,3,4,5	1,2a,3,5,10
B	1	2b,4,6,11,15,17,20,22,25,26,27,28,29,30,31,33,34,35,36,41,44,47
C	1,2,6,7	9,12,13,14,18,32,40,42,43,50

Type A CBMs – surface-binding CBMs

Type A CBMs include members of CBM families 1, 2a, 3, 5 and 10 that bind to insoluble, highly crystalline cellulose and/or chitin and mannan and show little or no affinity for soluble carbohydrates. These CBMs present a flat or platform-like binding surface, which is complementar to the flat surfaces presented by cellulose or chitin crystals (Tormo *et al.*, 1996). Type A CBMs show little or no affinity for soluble carbohydrates and the interaction of type A modules with crystalline cellulose is associated with positive entropy, demonstrating that the thermodynamic forces that drive the binding of CBMs to crystalline ligands are relatively unique among carbohydrate-binding proteins (Creagh *et al.*, 1996).

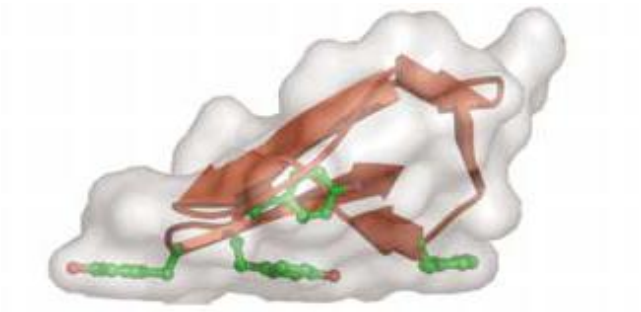


Figure 1.10 Example of a CBM belonging to the functional Type A.
Type A CBM1 (PDB code 1cbh), belonging to chitosanase from *Hypocrea jecorina L27*, bind via a flat planar yielded larger heats.

Type B CBMs – glycan-chain-binding CBMs

Type B CBMs bind to soluble polysaccharide chains and this type of CBMs, which currently includes examples from families 2b, 4, 6, 15, 17, 20, 22, 27, 28, 29, 30, 34, 36 and 44, have clearly evolved binding site topographies that are equipped to interact with individual glycan chains rather than to the crystalline surface of carbohydrates. The carbohydrate-binding sites of Type B CBMs are extended and are usually described as grooves or clefts comprising several subsites able to accommodate the individual sugar units of the polymeric ligand (Boraston *et al*., 2004). In general, the binding proficiency of this class of CBM is determined by the degree of polymerization of the carbohydrate ligand; biochemical studies frequently demonstrate increased affinities up to hexasaccharides and negligible interaction with oligosaccharides with a degree of polymerization of three or less (Boraston *et al*., 2004). Therefore, these CBMs

are usually described as 'chain binders'. The depth of these binding sites varies from very shallow to being able to accommodate the entire width of a pyranose ring.

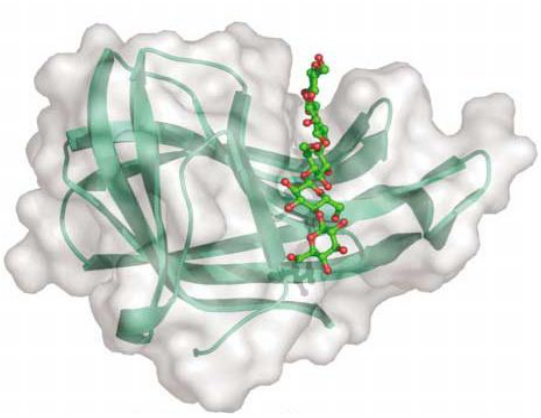


Figure 1.11 Example of a CBM belonging to the functional Type B.
The figure represents a cellopentaose molecule occupying a deep binding groove of CBM4-2 (PDB code 1GU3), which belongs to endo-1,4-glucanase from *Cellulomonas fimi ATCC 484*.

Type C CBMs – Small-sugar-binding CBMs

This unique class of CBMs have the lectin-like property of binding optimally to mono-, di- or tri-saccharides, thus lacking the extended binding-site grooves of Type B CBMs (Boraston *et al*., 2004). It should be emphasized, however, that the distinction between Type B and Type C CBMs can be subtle. For example, the Type B CBM6 module of the *Clostridium stercorarium* xylanase has a very similar fold to the Type C lectin-like CBM32 family, but apparently binds longer oligosaccharide ligands (Boraston *et al*., 2004). Nevertheless, it is apparent that the hydrogen-bonding network between protein and ligand is more extensive in Type C CBMs than Type B modules, consistent with their lectin-like properties (Boraston *et al*., 2004).

The Type C CBMs currently includes examples from families 9, 13, 14, 18, 32, 40, 42, 43 and 50 (Table 1.3). Members of families 13 (e.g. ricin toxin B-chain), 14 (e.g. tachycitin) and 18 (e.g. WGA) were first discovered as lectins with small-sugar-binding activity and have only subsequently been included as CBMs due to their discovery in a number of glycoside hydrolases that degrade plant structural carbohydrates (Boraston *et al*., 2004).

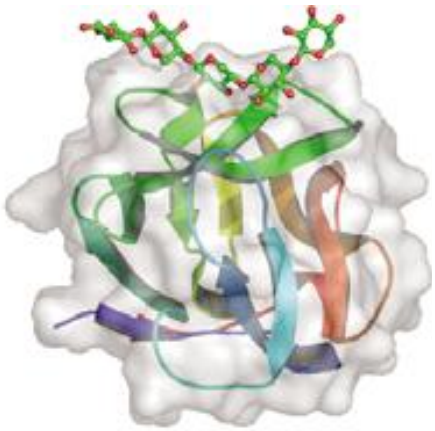


Figure 1.12 Example of a CBM belonging to the functional Type C.
The figure represents a xylopentaose molecule occupying the binding cleft of CBM13 (PDB code 1MC9), from xylanase from *Streptomyces lividans.*

Identification and characterization of Type C CBMs is lagging behind Type A and B CBMs, probably due to their limited presence in plant cell wall active glycoside hydrolases. Rather, Type C CBMs, particularly CBMs from families 13 and 32, appear to be more prevalent in bacterial toxins or enzymes (glycoside hydrolases and glycosyl transferases) that attack eukaryotic cell surfaces or matrix glycans (Boraston *et al.*, 2004).

1.2.1.3.2. Biological role of CBM.

Enzyme targeting

CBMs that bind to the surfaces of crystalline polysaccharides (referred to as Type A modules) can be appended to a variety of glycoside hydrolases. In contrast, CBMs that interact with single polysaccharide chains (Type B or C) bind to polysaccharides that are the substrates for the cognate catalytic module of the enzyme. For example, cellulases, xylanases and mannanases contain Type B CBMs that bind to cellulose, xylan and mannan, respectively. Thus the CBM maintains proximity of the appended catalytic domain with its target substrate within complex macromolecular structures, such as the plant cell wall. It is apparent that this targeting function is even more subtle than the somewhat crude partitioning of enzymes to the different polysaccharides of plant cell walls.

Although the interaction of CBMs with cellulose is occasionally irreversible, their contact with the cellulose surface is a dynamic process (Shoseyov *et al.*, 2006).

Jervis *et al.* using fluorescence recovery techniques with CBMs labeled with fluorescent tags, confirmed that CBMs from *Cellulomonas fimi* are mobile on the surface of crystalline cellulose (Jervis *et al.*, 1997). The mobility of CBMs may explain the function of CBMs other than to concentrate the catalytic activity on the substrate. This dynamic binding behavior of CBMs has a functional importance as often they are part of cellulases acting processively (Lehtiö *et al.* 2003). Although CBMs have a high affinity for carbohydrates, it is obvious that an irreversible binding would be fatal for the catalytic efficiency of the appended hydrolic domains.

Ligand binding and specificity: the role of aromatic residues.

The interaction of aromatic amino acid side chains with ligand is ubiquitous to CBM carbohydrate recognition. However, it has been demonstrated that binding-site topography is a key determinant of binding specificity (Boraston *et al*., 2004). The two major factors appear to be the location of aromatic amino acid side chains and loop structures that shape the binding sites to mirror the conformation of the ligand (Boraston *et al*., 2004). The side chains of tryptophan, tyrosine and, less commonly, phenylalanine form the hydrophobic platforms in CBM-binding sites, which can be planar, twisted or form a sandwich (Figure 1.13) (Boraston *et al*., 2004). As previously referred, Type A CBMs binds to crystalline cellulose through a flat platform containing aromatic residues (Figure 1.13A), most often tyrosines and tryptophans, separated by a distance corresponding to the length of the repeating unit, in cellulose 10,3 Å. In this case, the flat aromatic ring interacts with the pyranose rings of the polysaccharides (Tomme *et al.*, 1995). This driven interaction may be supplemented by few hydrogen bonds mediated by polar residues located at the binding interface (Tomme *et al.*, 1995). In contrast, type B CBMs binding sites are equipped to interact with individual glycan chains rather than crystalline surfaces (Boraston *et al*., 2004). In the binding sites of families 2b, 15, 17, 27, 29, 34 and 36, the apolar platform can be 'twisted' due to the rotation of the planes of two to three aromatic amino acid side chains relative to one another (Figure 1.13B) (Boraston *et al*., 2004). In type B, the binding cleft can also present a sandwich form, where the aromatic amino acid side chains sandwich a sugar unit in the ligand by stacking against the β and α face of the pyranose ring (Figure 1.13C). This latest case is common to family 4, 6, 9 and 22 CBMs. The sandwich and twisted platforms may be used concurrently in the same CBM and can both

accommodate the conformations of soluble oligosaccharide ligands (Boraston *et al*., 2004). CBMs appear to have preformed carbohydrate-recognition sites which mirror the solution conformations of their target ligands, thereby minimizing the energy required for binding (Boraston *et al*., 2004).

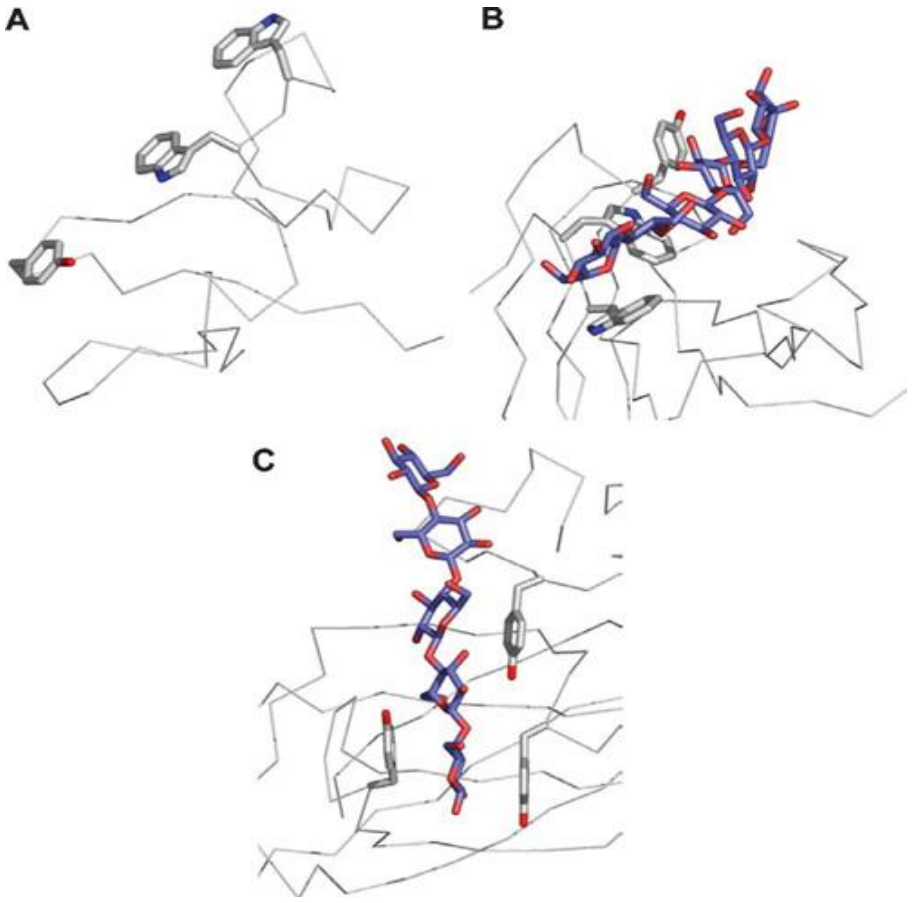


Figure 1.13 The three types of binding-site 'platforms' formed by aromatic amino acid residues. **(A)** The 'planar' platform in the family 10 Type A CBM, *Cj*CBM10. **(B)** The 'twisted' platform of the Type B family 29 CBM, *Pe*CBM29-2. **(C)** The 'sandwich' platform of the Type B family 4 CBM, *Cf*CBM4-2. The C-*α* backbone is shown as grey cylinders, the aromatic amino acid side chains forming the binding sites are shown in grey 'liquorice' representations, and the bound oligosaccharides are shown in blue 'liquorice' representation. Adapted from Boraston *et al*. (2004).

As with type A, in type B CBMs the aromatic residues play a pivotal role in ligand recognition and the orientation of these amino acids are key determinants of the CBMs binding specificity. For example, family 2 CBMs contain Type A and Type B members, named CBM2a and CBM2b, which bind crystalline cellulose or xylan, respectively. The explanation to this contrasting ligand specificity results from the rotation by 90º of the side-chain of one of the surface tryptophan (Trp^{259}) involved in the protein-carbohydrate interaction in CBM2b compared with its position in CBM2a

(Simpson *et al.* 2000). To prove this possibility, Simpson *et al.* (2000) showed that the ligand specificity of CBM2 is determined largely by a single amino acid, which controls the orientation of one of the tryptophan residues that interacts with the saccharide ligand. When the tryptophans are coplanar, the CBM recognizes the planar chains of cellulose, whereas when they are twisted into a near perpendicular arrangement, the protein recognizes the helical structure of xylan. Thus, in this CBM family, ligand specificity is determined largely by recognition of the three-dimensional shape of the polysaccharide ligand, rather than by specific hydrogen bonding patterns, as is typically seen in proteins that recognize monosaccharides (Simpson *et al.* 2000).

In contrast with Type A CBMs, direct hydrogen bonds may play a key role in defining the affinity and ligand specificity of type B CBMs, usually described as chain binders (Notenboom *et al.*, 2001). There is no evidence, however, indicating that water-mediated hydrogen bonds are critical in the interaction of CBMs with their target ligands (Pell *et al.*, 2003). Although the orientation and positioning of the aromatic residues in the binding sites of CBMs is the primary driver of specificity and affinity in these proteins, other interactions, including direct hydrogen bonds and calcium-mediated co-ordination, play a significant role in CBM ligand recognition (Boraston *et al*., 2004).

Non-hydrolytic substrate disruption and surface/interfacial modifications

It has been proposed that some CBMs, mainly of Type A, possess additional functions, such as the capacity for substrate disruption. The initial evidence supporting this property was described for the N-terminal family 2a CBM of Cel6A from *Cellulomonas fimi* (Boraston *et al*., 2004), which appears to mediate non-catalytic disruption of the crystalline structure of cellulose, resulting in the release of small particles without any detectable hydrolytic activity. In addition, this disruptive effect enhanced the degradative capacity of the catalytic module (Boraston *et al*., 2004).

Other studies indicated that treatment of cellulose fibers with CBMs alters their interfacial properties (Suurnäkki *et al.*, 2000; Pala *et al.*, 2001). It has also been demonstrated that CBMs prevent the flocculation of crystalline cellulose (Shoseyov *et al.* 2006), while Lee et al. (2000) provided the first physical evidence that CBM from *Trichoderma reesei* caused peeling and smoothed surface on cotton fibers. This non-

disruptive ability provides CBMs with the potential for usage in numerous biotechnological applications, such as in the textile and paper industries.

1.2.1.3.3. Multivalency of CBMs

The weak interaction between a carbohydrate-binding protein and its ligand are often compensated in nature by multivalent interactions. In these cases, multiple clustered carbohydrate-binding sites interact simultaneously with carbohydrate ligands, which present multiple recognition elements, resulting in increased association constants relative to any one of the isolated protein–carbohydrate interactions (Boraston *et al*., 2004). It is clear that clustered carbohydrate-binding sites can result from (i) a single protein having multiple binding sites, (ii) the association of two or (iii) more univalent carbohydrate-binding proteins into multivalent quaternary structures, or clustering of receptors, for example, in cell membranes.

To date, no CBM has been found to form quaternary structures in its natural state but multiple CBMs are found frequently in glycoside hydrolases, which effectively become multivalent carbohydrate-binding proteins and overcoming relatively weak binding (Boraston *et al*., 2004). The presence of multiple CBMs in glycoside hydrolases appears to occur most frequently in thermophilic or hyperthermophilic enzymes as a possible response to the need for these proteins to overcome the loss of binding affinity that accompanies most molecular interactions at elevated temperatures (Boraston *et al*., 2003a).

The first CBMs in tandem to be investigated were the family 2b CBMs of *Cellulomonas fimi* xylanase 11A (Bolam *et al.*, 2001). While the individual association constants for xylan were of approximately 10^4 M^{-1}, CBMs linked in tandem displayed an association constant of approximately 10^6 M^{-1}. The same was observed for the three family 6 CBMs of the *Clostridium stercorarium* putative xylanase (Boraston *et al.*, 2002b) and the tandem CBM17 and CBM28 modules from *Bacillus* sp. 1139 Cel5 (Boraston *et al.*, 2003a).

Individual CBM modules containing multiple carbohydrate-binding sites occur in a variety of CBM families. This was first proposed for the family 13 Type C CBM from *S. lividans* xylanase 10A it was subsequently demonstrated by mutagenesis, NMR and X-ray crystallography that this module had three separate binding sites, one in each of its trefoil subdomains , Figure 1.14a (Boraston *et al.*, 2000b; Notenboom *et al.*, 2002). The family 6 CBM from *Cellvibrio mixtus* endoglucanase 5A has two binding

sites (Figure 1.14b), named 'cleftA' the one that can accommodate the chain ends of *β*-1,4-glucans, *β*-1,3-glucans and xylans and 'Cleft B' that binds to internal regions of *β*-1,4-glucans and mixed *β*-(1,4)(1,3)-glucans (Pires *et al*., 2004; Henshaw *et al.*, 2004).

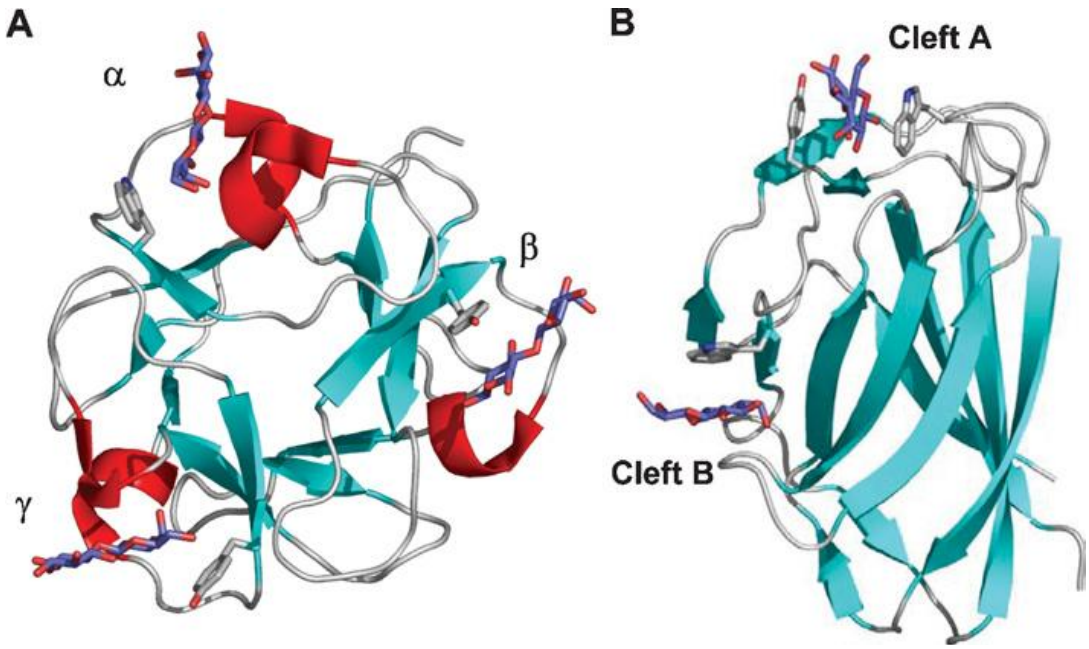


Figure 1.14 Multivalent CBMs.
The family 13 CBM from S. lividans xylanase 10A, SlCBM13, is shown in complex with three lactose molecules occupying its three binding sites (**A**), while the family 6 CBM from Cellvibrio mixtus endoglucanase 5A, CmCBM6, is shown in complex with two cellobiose molecules occupying its two binding sites (**B**). Both are shown as ribbon diagrams with relevant aromatic amino acid side chains in the binding sites shown in grey 'liquorice' representation. The three *β*-trefoil subdomains of SlCBM13 are labelled as *α*, *β* and *γ* , and the two binding sites of CmCBM6 are labeled according to the cleft in which they reside. Adapted from Boraston *et al*. (2004).

1.3. OBJECTIVES OF THIS WORK

The work presented here aims to elucidate several unresolved questions about the structure and function of CEs and CBMs from *Clostridium thermocellum* cellulosome. Specifically, the aims of this project can be summarized as follows:

1. To determine the structure and biochemical properties of a family 3 carbohydrate esterase from *Clostridium thermocellum* cellulosome, *Ct*Ces3-1 (Chapter 2);

2. To identify the role of CBM35 in pectin degradation by *C. thermocellum* cellulossome (Chapter 3);

3. To reveal the biochemical properties and ligand specificity of a family 6 CBM from the *Cellvibrio mixtus* multi-modular glycoside hydrolase, *Cm*Cel5A (Chapter 4);

4. To characterize the function of *Ct*CBM55, the founder member of a novel CBM family which binds complex polysaccharides decorated with galactose (Chapter 5).

CHAPTER 2 Crystal structure of a cellulosomal family 3 carbohydrate esterase from clostridium thermocellum provides insights into the mechanism of substrate recognition.

Márcia A. S. Correia[1], José A. M. Prates1, Joana Brás[1], Carlos M. G. A. Fontes[1], Joseph A. Newman[2], Richard J. Lewis[2], Harry J. Gilbert[2] and James E. Flint[2]

[1]CIISA - Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [2]Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, UK

Adapted from *Journal of Molecular Biology* (2008) 379, 64–72

Abstract

The microbial degradation of the plant cell wall is of increasing industrial significance, exemplified by the interest in generating biofuels from plant cell walls. The majority of plant cell wall polysaccharides are acetylated, and removal of the acetyl groups through the action of carbohydrate esterases greatly increases the efficiency of polysaccharide saccharification. Enzymes in carbohydrate esterase family 3 (CE3) are common in plant cell wall degrading microorganisms but there is a paucity of structural and biochemical information on these biocatalysts. *Clostridium thermocellum* contains a single CE3 enzyme, *Ct*Ces3, which comprises two highly homologous (97% sequence identity) catalytic modules appended to a C-terminal type I dockerin that targets the esterase into the cellulosome, a large protein complex that catalyses plant cell wall degradation. Here, we report the crystal structure and biochemical properties of the N-terminal catalytic module (*Ct*Ces3-1) of *Ct*Ces3. The enzyme is a thermostable acetyl-specific esterase that exhibits a strong preference for acetylated xylan. *Ct*Ces3-1 displays an α/β hydrolase fold that contains a central five-stranded parallel twisted β-sheet flanked by six α-helices. In addition, the enzyme contains a canonical catalytic triad in which Ser44 is the nucleophile, His208 is the acid base and Asp205 modulates the basic nature of the histidine. The acetate moiety is accommodated in a hydrophobic pocket and the negative charge of the tetrahedral transition state is stabilized through hydrogen bonds with the backbone N of Ser44 and Gly95 and the side-chain amide of Asn124.

2.1. INTRODUCTION

The plant cell wall, which comprises the most abundant source of organic carbon in the biosphere with an estimated 10^{11} tons recycled annually, is a highly complex composite macromolecular structure consisting largely of carbohydrate polymers (Brett *et al.*, 1996). The microbial degradation of the plant cell wall is an important biological process, and is of increasing industrial significance exemplified by the interest in second-generation biofuels derived from lignocellulosic biomass (Ragauskas *et al.*, 2006; Sticklen *et al.*, 2006). The majority of the matrix plant cell wall polysaccharides, including xylan, mannan, glucomannan and pectin, can be acetylated and, in some instances, ferulyated, which sterically restricts the access of the repertoire of glycoside hydrolases that attack the glycosidic linkages in these polymers (Brett *et al.*, 1996). Removal of the acetyl groups though the action of carbohydrate esterases greatly increases the efficiency of polysaccharide saccharification through the action of glycoside hydrolases (Ferreira *et al.*, 1993). Thus, acetyl esterases are important components of the plant cell wall-degrading apparatus of saprophytic microorganisms.

Enzymes that degrade complex carbohydrates are grouped into sequence-based families.Within a given family, the overall fold, catalytic mechanism and the catalytic apparatus are completely conserved (Henrissat *et al.*, 1991). Currently there are 15 carbohydrate esterase families (CEs) and there is some three-dimensional structural information available for at least one member of families CE1, CE4, CE5, CE6, CE7, CE8, CE9, CE10, CE11, CE13 and CE14. Apart from CE4, all representatives of these families display a typical α/β hydrolase fold. The feruloyl and acetyl esterases from families CE1, CE5, CE6 and CE7, which function as typical serine esterases, contain classical Ser-His-Asp catalytic triads. By contrast, members of CE4 exhibit a greatly distorted $(\beta/\alpha)_8$ barrel and the activity of at least some members of this family is metal ion dependent (Taylor *et al.*, 2006). The diversity of esterase specificities displayed by members of the various CEs reflects intrinsic differences in the structure and surface topography of the catalytic sites of these enzymes. *Clostridium thermocellum* is an anaerobic thermophilic bacterium that, through the action of a supramolecular multi-protein complex of cellulases and hemicellulases termed the cellulosome, efficiently degrades plant cell wall polysaccharides (Bayer *et al.*, 2004). The assembly of the extracellular complex results from the interaction of "type I dockerin" modules located in the catalytic subunits with the multiple "cohesin" modules of the scaffolding protein. Inspection of the genome sequence of *C. thermocellum* revealed 71 polypeptides

containing type I dockerins, nine of which possess carbohydrate esterase modules belonging to CE1, CE2, CE3, CE4, CE6, CE8 and CE12. Interestingly, some of these catalytic modules are found in bi-functional enzymes containing GH10 and GH11 xylanase or GH5 endoglucanase modules, reinforcing the importance of the synergistic interaction between carbohydrate esterases and glycoside hydrolases that attack, particularly, the backbone of xylan, a major hemicellulosic polysaccharide. To explore the mechanism by which the cellulosome deacetylates polysaccharides we have characterized the N-terminal catalytic module of the single CE3 esterase, referred to here as *Ct*Ces3-1, present in the plant cell wall-degrading cellulosome complex. The enzyme is a carbohydrate esterase that deacetylates acetylated xylan. The protein displays a typical α/β hydrolase fold and contains the classic Ser-His-Asp triad of serine esterases. Site-directed mutagenesis showed that Ser44 is the nucleophile, His208 activates the catalytic serine by abstracting a proton, and Asp205 increases the basic character of the histidine.

2.2. MATERIALS AND METHODS

2.2.1. Gene cloning and protein expression

The region of the *C. thermocellum* esterase gene, *ces3*, encoding the N-terminal CE3 esterase module, *Ct*Ces3-1 (residues 34–237 of the full-length enzyme), was amplified by PCR from genomic DNA (strain ATCC 27405) using the thermostable DNA polymerase Pfu Turbo (Stratagene) and the primers listed in Table 2.1 that contain *Nhe*I and *Xho*I restriction sites, respectively, depicted in bold. The DNA product was cloned into the *Nhe*I and *Xho*I sites of the *Escherichia coli* expression vector pET21a (Novagen) to generate pCE3. *Ct*Ces3-1 encoded by pCE3 contains a C-terminal His_6 tag.

Table 2.1 Primers used for the cloning and mutagenesis of CBM30 and CBM44.
Enzyme and restriction site are depicted in bold.

Construct	Primers	Sequence (5'-3')
pCE3	Forward	CTC**GCTAGC**AAAACGATAAAAATCATGCC
	Reverse	CAC**CTCGAG**CTGTGTTTCTCCAGCCAAAG
S44A	Forward	CCTGTCGGAGAT**GCC**TGCACCGAAGG
	Reverse	CCTTCGGTGCA**GGC**ATCTCCGACAGG
D205A	Forward	CGATATTTCATGG**GCG**GGTTTGCACTTGAGC
	Reverse	GCTCAAGTGCAAACC**CGC**CCATGAAATATCG
D205N	Forward	CGATATTTCATGG**AAT**GGTTTGCACTTGAGC
	Reverse	GCTCAAGTGCAAACC**ATT**CCATGAAATATCG
H208A	Forward	CATGGGATGGTTTG**GCC**TTGAGCGAAATAGG
	Reverse	CCTATTTCGCTCAA**GGC**CAAACCATCCCATG

2.2.2. Protein expression and purification

E. coli BL21 cells harbouring pCF1 were cultured in Luria-Bertani broth at 37°C to mid-exponential phase (A_{600} nm 0.6) and recombinant protein expression was induced by the addition of isopropyl 1-thio-β-d-galactopyranoside (IPTG) to 1 mM final concentration and incubation for a further 5 h at 37 °C. *Ct*Ces3-1 was purified by immobilized metal ion affinity chromatography (IMAC) using Talon™ resin and elution in 20 mM Tris–HCl buffer (pH 8.0) containing 300 mM NaCl and 100 mM imidazole. The eluted esterase was then dialyzed against buffer A (20 mM Tris–HCl, pH 8.0) and applied to a Bio-Rad Q12 column. *Ct*Ces3-1 was eluted with a 400 ml 0–500 mM NaCl gradient in Buffer A. The fractions containing esterase activity were concentrated using a 10 kDa cutoff Vivaspin 20 centrifugal concentrator and applied to a Superdex™ 200 26/60 Hiload™ column (Amersham) equilibrated in 10 mM Tris-HCl buffer (pH 8.0), 150 mM NaCl. The protein was eluted at a flow-rate of 1 ml min−1. Both chromatography steps employed a Bio-Rad FPLC system. Purified *Ct*Ces3-1 was judged homogeneous by SDS-PAGE. The same protocol was used to produce SeMet-containing *Ct*Ces3-1, except that the enzyme was expressed in *E. coli B834* (Novagen) using growth conditions as described (Ramakrishnan *et al.*, 1993), and 2 mM 2-mercaptoethanol was included in all buffers during purification up to the point of IMAC. The concentration of 2-mercaptoethanol was raised to 10 mM for all subsequent steps. The purified SeMet-containing enzyme eluted from the gel-filtration chromatography column was exchanged into doubledistilled water containing 10 mM DTT using an Amersham PD10 column.

2.2.3. Mutagenesis

Site-directed mutagenesis was carried out employing a PCR-based QuikChange site-directed mutagenesis kit (Stratagene) according to the manufacturer's instructions, using pCF1 as the template and the primers used to generate these mutants are displayed in Table 2.1, with the mutagenic site in bold face. The mutated DNA sequences were sequenced to ensure that only the appropriate mutations had been incorporated into the nucleic acid.

2.2.4. Enzyme assays

Substrates used in the enzyme assays described below were purchased from Sigma, except acetylated Konjac glucomannan, which was purchased from Megazyme International, and acetylated xylan, which was prepared from birchwood xylan (Sigma) as described previous (Johnson *et al.*, 1988). To screen for activity against acryl acetates, 1 ml enzyme reactions were carried out in 50 mM sodium phosphate buffer (pH 7.0) containing 1 mg ml^{-1} of BSA and 1 mM substrate. The reaction was initiated by the addition of 12 nM enzyme and the product release was monitored at 400 nm (4-nitrophenolate) or terminated with an equal volume of 10% (w/v) SDS containing 0.01% (w/v) Fast garnet GBC (Sigma), and the A560 nm was determined (Napthyl) after incubation at room temperature for 15 min. To determine the rate of deacetylation of acetylated polysaccharides, nucleotides and cephalosporin, reactions were carried out as described above except that the release of acetate was determined by HPLC using an isocratic flow-rate at 1.5 ml min^{-1} of 15 mM H_2SO_4 on a Bio-Rad Aminex HPX-87H ion-exclusion column.

2.2.5. Crystallization, data collection, structure solution and refinement

SeMet crystals of *Ct*Ces3-1 were grown by vapourphase diffusion using the hanging-drop method with an equal volume (1 μl) of protein (40 g l−1 in 10mM DTT) and reservoir solution (2% (w/v) PEG 400, 0.1 M NaHepes, pH 7.5, and 2.0 M ammonium sulphate). Similarly, native crystals were grown in the presence of acetate, with the exception that the protein solution contained 100 mM sodium acetate and no DTT. Crystals grew over a period of one week and were cryo-protected by the addition of glycerol to the crystallization mother liquor to a final concentration of 20% (v/v). The structure of SeMet data was determined using the SAD method at a wavelength optimised for the f ″ signal of the selenium. SeMet SAD data were collected at SRS Daresbury on beamline 10.1 to a resolution of 1.4 Å. A total of 180 images were collected at a wavelength of 0.933 Å. The data were recorded on an ADSC Q4 CCD detector at an oscillation angle of 1° to a resolution of 1.6 Å followed by a further 180 images at the same oscillation angle to a maximum resolution of 1.4 Å. The two sets of data were merged and scaled using SCALA (Evans *et al.*, 1993). Three Se sites were found using the HKL2MAP (Pape *et al.*, 2004) interface of SHELXD (Shneider *et al.*, 2002). Heavy-atom phasing was performed with MLPHARE (CCP4, 1994), and these

intial phases were improved by solvent flattening in DM (Cowtan *et al.*, 1994). Automated model building and refinement using ARP/wARP (Murshudov *et al.*, 1997) and REFMAC (Lamzin *et al.*, 1997) resulted in a model with a crystallographic R-factor of 0.145% (Rfree 0.168%), which was completed with manual correction using COOT (Emsley *et al.*, 2004), and refinement with REFMAC (Lamzin *et al.*, 1997). The native structure was solved using MOLREP (Vagin *et al.*, 1997) with SeMet *Ct*Ces3-1 as the search model.

2.2.6. Protein Data Bank accession code

The coordinates and observed structure factor amplitudes have been deposited in the Protein Data Bank with accession code 2VPT.

2.3. RESULTS AND DISCUSSION

2.3.1. Biochemical properties of *Ct*Ces3

The genome of *C. thermocellum* contains a single CE3 esterase (*Ct*Ces3; Cthe_0798) gene that is located immediately upstream of *cel5C*, which directs the synthesis of a modular enzyme comprising a GH5 endoglucanase catalytic module, a type I dockerin, and a C-terminal CE2 esterase. *Ct*Ces3 comprises 528 amino acids and consists of a signal peptide followed by two tandemly repeated CE3 modules that are 97% identical, designated *Ct*Ces3-1 and *Ct*Ces3-2, and a C-terminal type I dockerin module (Figure 2.1). DNA encoding the N-terminal CE3 module was cloned into the *E. coli* expression vector pET21a to generate pCF1 as described in Materials and Methods. *E. coli* BL21 harbouring pCF1 expressed large amounts of soluble *Ct*Ces3-1, which was purified to electrophoretic homogeneity. The biochemical properties of *Ct*Ces3-1 were assessed using both natural and artificial substrates. The reaction conditions employed in these assays are described in Materials and Methods. The enzyme did not attack aryl-ferulates or aryl-coumorates but did hydrolyse a range of acetate-containing substrates with good leaving groups, displaying a specific activity against α-napthyl acetate of 45.2 U mg^{-1} protein (1 U is the hydrolysis of 1 μmol of substrate per minute). These data indicate that *Ct*Ces3-1 is an acetyl esterase. The k_{cat}/K_m for *Ct*Ces3-1 hydrolysis of 4-ρ-nitrophenyl acetate (PNPAc) is 152 $min^{-1}M^{-1}$; k_{cat} and K_m could not be determined individually because there was a direct relationship between rate and substrate concentration up to 10 mM PNPAc (maximum concentration of soluble substrate).

While the enzyme displayed no detectable activity against acetylated glucomannan, the esterase did deacetylate acetylated birchwood xylan, displaying a specific activity of 70 U mg^{-1} protein. The enzyme was able to deaceylate oligosaccharides derived from acetylated xylan by a GH11 xylanase; however, *Ct*Ces3-1 exhibited no activity against glucomannooligosaccharides. The inability of *Ct*Ces3-1 to hydrolyse other representative acetylated compounds, Cephalosporin C, N-acetyl-glucosamine and acetyl deoxyadenosine, suggests that the enzyme displays specificity for acetylated xylans, which likely comprises the natural substrate, although the esterase can hydrolyse synthetic molecules in which the acetyl moiety is esterified to a good leaving group. The high K_m for PNPAc likely reflects the lack of affinity for the aromatic phenol, as the enzyme displays a high level of specificity for xylan-based molecules. Thermostability studies showed that *Ct*Ces3-1 retains maximum catalytic activity when incubated for 15 min up to 70 °C, after which the esterase is rapidly inactivated displaying no activity after treatment at 74 °C for 15 min. The pH optimum for *Ct*Ces3-1 was pH 7.0 and thus the biophysical properties of the esterase are entirely consistent with the observation that *C. thermocellum* is a thermophile that displays a maximum growth rate at neutral pH (Bayer *et al.*, 1983). Within the 36 members of the CE3 family, the biochemical properties of only three enzymes, apart from *Ct*Ces3-1, have been investigated. The two from *Ruminococcus flavafaciens* (XynB and CesA) and one from *Neocallimastix patriciarum* (BnaC) all hydro lysed α/β-napthyl acetate and PNPAc but displayed reduced or no activity against esters of longer-chain aliphatic carboxylic acids (Dalrymple *et al.*, 1997; Aurilia *et al.*, 2000). These data indicate that these enzymes are not lipases, or utilise aromatic acid esters such as ferulate as substrates, eliminating the possibility that they display cinnamoyl esterase activity (Dalrymple *et al.*, 1997; Aurilia *et al.*, 2000). Although all three enzymes hydrolysed acetylated xylans, their capacity to attack other acetylated plant structural polysaccharides such as galactomannan or galactoglucomannan was not reported (Dalrymple *et al.*, 1997; Aurilia *et al.*, 2000). Here, we show that *Ct*Ces3-1 is able to attack acetylated xylan but not other esterified polysaccharides, which, in conjunction with the studies on the other CE3 members (Dalrymple *et al.*, 1997; Aurilia *et al.*, 2000), indicates that (natural) substrate specificity in this family is restricted to xylans. The substrate specificity of enzymes in other CEs is variable. For example, the *Bacillus subtilis* CE7 esterase described by Vincent and colleagues (Vincent et al., 2003) displays a preference for small acetylated substrates but not for polysaccharides.

However, unlike *Ct*CE3, the CE7 enzyme, apart from a size limitation, exhibits little specificity for the alcohol component of the substrate, as it is active on acetylated xylooligosaccharides, deoxyadenosine and Cephalosporin C (Vincent *et al.*, 2003). The preference of the *B. subtilis* enzyme for small molecules is related to the steric effects imposed by its hexameric structure. By contrast, CE4 esterases display absolute specificity for either chitin or acetylated xylan (Blair *et al.*, 2006; Taylor *et al.*, 2006), while a CE1 enzyme from *C. japonicus* exhibits little specificity for the alcohol or carboxylic acid component of the substrate (Ferreira *et al*., 1993).

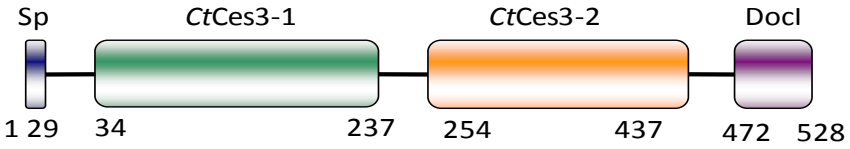


Figure 2.1 Molecular architecture of CtCE3 from *Clostridium thermocellum*.
Domain bondaries are numbered with amino acid sequence and the abbreviated modules are as follows: Sp, signal peptide; DocI, type I dockerin modules.

2.3.2. Crystal structure of *Ct*Ces3-1

The crystal structure of *Ct*Ces3-1 was solved using the single-wavelength anomalous dispersion (SAD) method employing selenomethionine-labelled protein and refined to a crystallographic R-factor of 0.145 (Rfree 0.168) using data extending to 1.4 Å resolution. The *Ct*Ces3-1 crystals belong to space group $P3_121$ with unit cell dimensions a=b=78.0 Å, c=66.8 Å. The X-ray diffraction data and structure refinement statistics are shown in Table 2.2. Overall, the electron density is of high quality throughout the entire length of the protein and the final model contains 201 amino acid residues (extending from 34–237 in the full-length sequence of *Ct*Ces3), 175 water molecules and one calcium ion. The structure generally has backbone torsion angles in the most favoured region (90.6% of the bonds), and none in the generously allowed or disallowed regions of the Ramachandran plot, according to PROCHECK (Laskowski *et al.*, 1993).

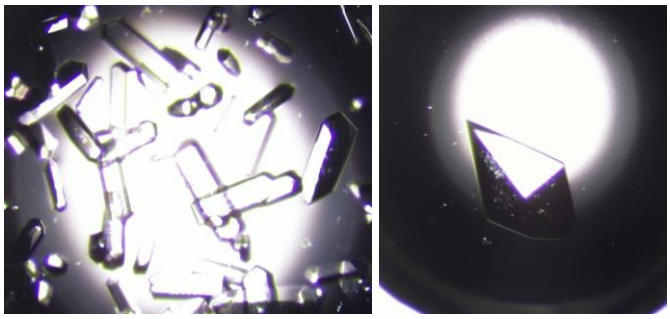


Figure 2.2 Crystals of Selenomethionine-*Ct*Ce3-1.
The Crystals were obtained by hanging-drop vapor diffusion in the presence of 2% PEG 400, 0,1M NaHepes pH 7,5 and 2,0 M ammonium sulphate

Table 2.2 X-ray diffraction collection and structure refinement of *Ct*CE3-1

Space group	$P3_121$
Unit cell parameters	
a (Å)	78.0
b (Å)	78.0
c (Å)	66.8
α (°)	90
β (°)	90
γ (°)	120
Matthew´s coefficient	2.64
Solvent content (%, v/v)	53
Resolution range (Å)	47.5-1.4 (1.44-1.40)
No. unique reflections	46,638 (6749)
Completeness (%)	99,9 (99,7)
Redundancy	10.9 (10.6)
$\mathbf{R_{merge}}$	0.09 (0.42)
I/(σ)	18.4 (4.5)
Refinement	
Rwork	0.145 (0.144)
Rfree	0,168 (0.18)
rsmd from ideal	
Bond lengths (Å)	0.026
Bond angles (°)	1.406
No. protein atoms	1585
No. water molecules	175
No. calcium ions	1
Mean B, all atoms (Å)	13.1

2.3.3. Overall structure

*Ct*Ces3-1, (Figure 2.3) displays an atypical α/β-hydrolase fold (also known as the SGNH-hydrolase superfamily fold), consisting of repeating β-α-β motifs that form a curved central five-stranded parallel β-sheet, in the strand order of β2, β1, β3, β4 and β5 with strand β2 being interrupted by a hairpin like loop. The sheet packs against two α-helices (α1 and α6) on the concave side and three α-helices (α2, α4 and α5) on the convex side, all of which are antiparallel to the β-strands. There is an additional α- helix (α3) between strands β-3 and β-4, and two small 310 helices between β1 and α1, and

between β5 and α6. In common with the majority of modules present in complex carbohydrate-modifying enzymes, the N and C termini are in close proximity, with the first and last residues in the model being separated by approximately 7 Å.

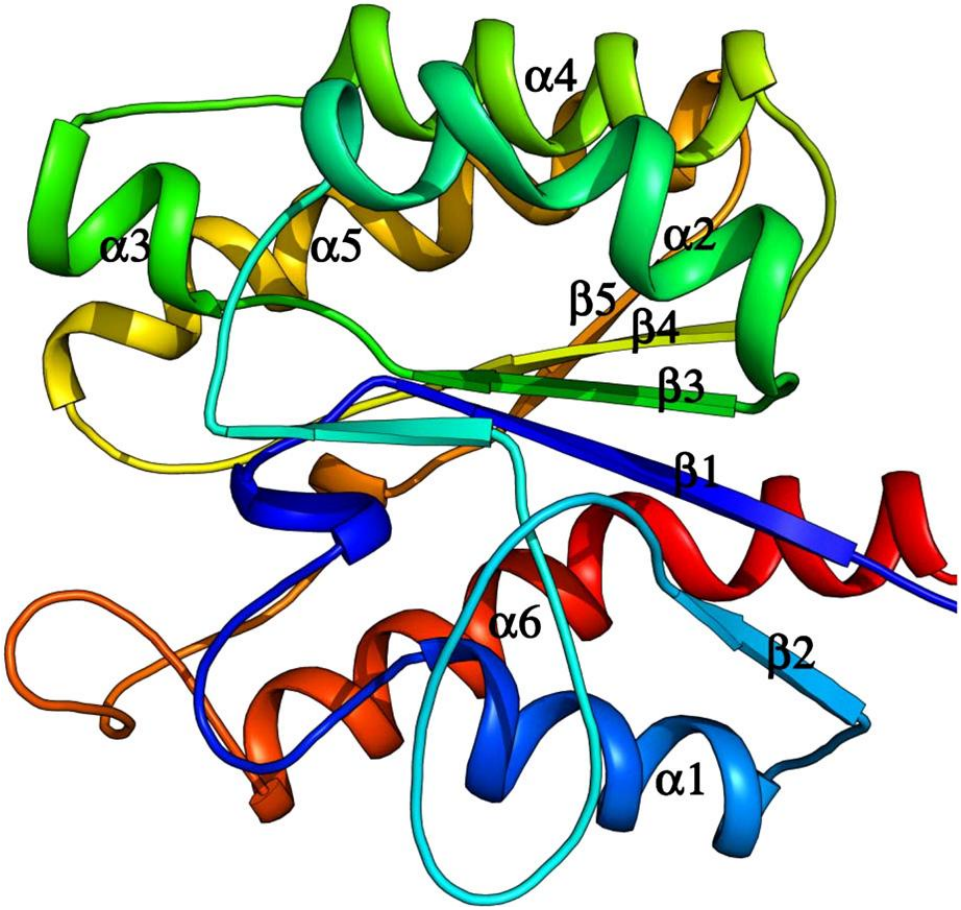


Figure 2.3 Structure of the *Ct*Ces3-1 monomer coloured as a blue to red gradient from the N terminus to the C terminus with key secondary structural elements labelled.
This and all subsequent molecular graphics images were prepared using the program PYMOL [http://pymol.sourceforge.net/].

Analysis of the crystallographic symmetry suggested a candidate dimer interface that could be generated by the 2-fold symmetry operator −x, −x+y, −z+1/3. This interface buries an area of 760 Å^2, which corresponds to 9% of the total monomer accessible surface area, which is within the typical range (6–20%) observed for homodimers (Jones *et al.*, 1995). The putative interface is formed by the β1–α1 loop, the hairpin within β2, and the β5–α6 loop. This interface is 69% polar in nature and contains six direct inter-molecular hydrogen bonds as calculated by the program HBPLUS using a distance cut-off of 4 Å (McDonald *et al.*¸1994). It is unclear whether this interface is biologically significant or merely a product of the crystallisation process, since *Ct*Ces3-1 is only a fragment of a larger multi-modular protein that may

oligomerise in an entirely different way. Alternatively, this surface may reflect the inter-module interfaces found in the full-length enzyme.

The $2F_{obs}–F_{calc}$ and $F_{obs}–F_{calc}$ electron density maps in the early stages of the model building and refinement process revealed a significant positive density feature that could not be attributed to a protein feature or a water molecule. This feature was assigned as a calcium ion on the basis of the pentagonal bipyramidal geometry of the co-ordinating ligands, and the σ level of the feature in the $2F_{obs}– F_{calc}$ electron density maps (17 σ) which, when compared to those of selenium (21 σ) and sulphur (10 σ) in the structure, is approximately consistent with the number of electrons in a calcium ion. The calcium ion is located directly above the N-terminal end of the central β-sheet, close to β-2 and is coordinated by the main-chain carbonyl of Asp73, Phe74, Asp87 (bidentate), Asp89 and two water molecules, all at distances between 2.3 Å and 2.5 Å (Figure 2.4). Treatment of the enzyme with EDTA did not inactivate the esterase; thus the calcium ion does not seem to be a component of the catalytic apparatus and is likely to have a structural role within the protein. Although this is consistent with the observation that metal ions do not play an important role in the enzyme activity of the majority of CEs, in at least one member of CE4 a divalent cobalt ion has a critical role in the catalytic mechanism of the esterase (Taylor *et al.*, 2006).

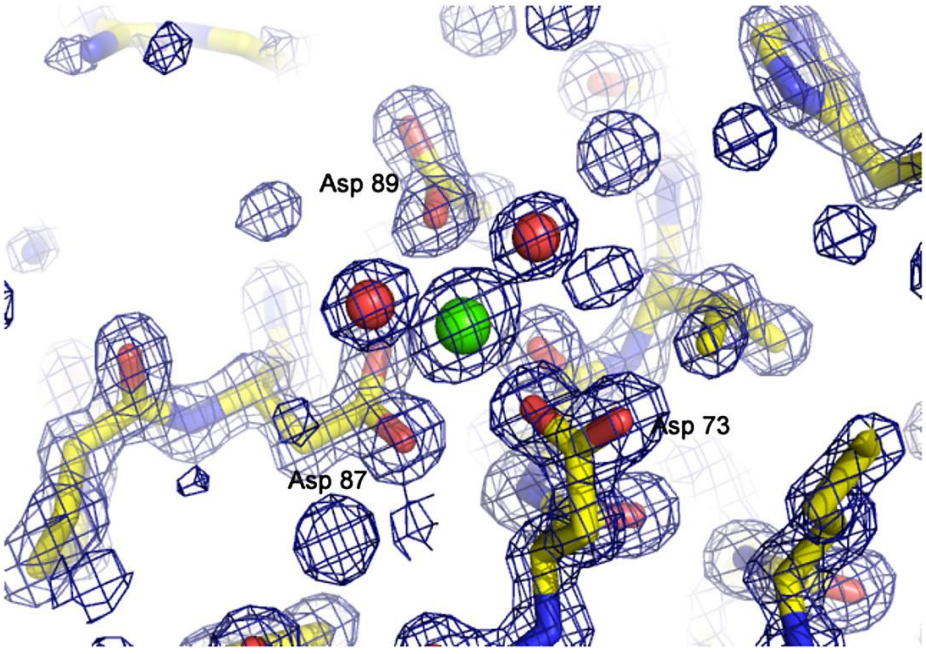


Figure 2.4 The 1.4 Å resolution 2Fo–1Fc electron density map contoured at 2σ covering the final refined coordinates of *Ct*Ces3-1 in the vicinity of the calcium ion-binding site.
Residues are shown in stick format and coloured according to atom type with carbon yellow, the calcium ion (shown as a green sphere) is coordinated by the side chain of three Asp residues, the main-chain carbonyl of Phe44 and two water molecules (shown as red spheres) in a pentagonal bipyramidal geometry.

2.3.4. Structural homologues

The co-ordinates of *Ct*Ces3-1 were submitted to a structure-based similarity search using SSM (Krissinel *et al.*, 2004) in order to determine the closest structural homologues. Overall, four unique hits were selected for comparison with *Ct*Ces3-1 with structural similarities significantly above the background level (Z-scores ranging from 7.5 to 6.4). The first two of these (PDB codes 1VJG and 1YZF) are products of structural genomics studies, are not characterised biochemically, and have been annotated as a putative lipase and a lipase/acylhydrolase, respectively. The third hit (1U8U) found by the SSM search is a thioesterase I/protease I/lysophospholipase L1 (TAP) from *E. coli*. (Lo *et al.*, 2003; Lo *et al.*, 2005) SSM also revealed a close match to the bovine platelet-activating factor acetylhydrolase (PAF-AH(Ib)) (1WAB). Not surprisingly, all of these enzymes belong to the α/β-hydrolase fold (SGNH hydrolase superfamily) and share a highly conserved core fold (Figure 2.5a), despite the low level of sequence similarity (~20% pairwise identities).

The enzymes of the α/β-hydrolase fold catalyse the hydrolysis of ester, thioester and amide bonds in a range of substrates and were first identified from sequence and structural analysis of the rhamnogalacturonan acetylesterase from Aspergillus aculeatus (Molgaard *et al.*, 2000). Other members belonging to this superfamily share sequence conservation over four separate blocks of residues (I–IV), of which the only strictly conserved residues from each motif were serine, glycine, asparagine, and histidine, respectively. As would be expected given their strict conservation, these residues are catalytically important. Currently, there is structural information for at least five distinct members of this family (1VJG, 1YZF and the structures described in Ho *et al.*, 1997; Molgaard *et al.*, 2000; Lo *et al.*, 2003), all of which share the same variant of the α/β-hydrolase fold that lacks the canonical nucleophile elbow. Instead, the catalytic serine is located on the loop connecting β1 and α1, and the β-sheet contains five rather than the normal eight strands. All these enzymes, however, contain the Ser/His/Asp catalytic triad in the active site.

2.3.5. *Ct*Ces3-1 active site

Comparing the structure of *Ct*Ces3-1 to other SGNH hydrolases has allowed the identification of the active site of the enzyme, which is located above the N-terminal end of the central β-sheet and close to β-strands 1, 2, 3, helix α3 and a number of loops.

Interestingly, part of this region extends to the putative dimer interface, indicating that this surface, if biologically significant, could contribute to shielding the active site from solvent. It is clear that the catalytic apparatus is retained in *Ct*Ces3-1 with respect to both the residues that form the catalytic triad (Ser44, Asp 205 and His208) and the amino acids that make hydrogen bonds with the transition state in the oxyanion hole (Ser44, Gly95 and Asn124). In all of these enzymes, the serine is believed to act as the catalytic nucleophile that attacks the substrate, whereas the histidine activates the nucleophile by abstracting a proton from Ser44 that it then donates to the alcohol leaving group. The aspartate stabilises the positive charge on the histidine by hydrogen bonding with the imidazole ring. In addition, the backbone amides of Ser44 and Gly95 and the $O^{\delta 1}$ of Asn124 serve as proton donors to the oxyanion hole that accommodates the negatively charged tetrahedral transition state. Inspection of the three dimensional structure of *Ct*Ces3-1 supports the proposed role of the catalytic triad. Thus, the O^{γ} of Ser44 makes a hydrogen bond with $N^{\varepsilon 2}$ of His208, while the $O^{\delta 2}$ of Asp205 is within hydrogen bonding distance (2.7 Å) of the histidine $N^{\delta 1}$ (Figure 2.5b).

In an attempt to investigate the mode of substrate binding and substrate specificity of this enzyme, the structure of *Ct*Ces3-1 was determined by co-crystallization in the presence of 100 mM sodium acetate, which unfortunately failed to yield any well defined electron density features in the proposed oxyanion hole. However, in the free enzyme structure the backbone amides of Ser44 and Gly95 and Oδ1 of Asn124 are all capable of making a hydrogen bond to a water molecule that is in a position believed to be occupied by the O− species of the tetrahedral transition state (Figure 2.5b), indicating that the oxyanion hole is present, but unoccupied. An acetate ion is located in the active site of PAF-AH(Ib),21 with the methyl group protruding into a hydrophobic pocket. The equivalent region in *Ct*Ces3-1 comprises a much deeper cavity (Figure 2.5b), possibly explaining how the esterase can accommodate an acetyl group appended to a more decorated polysaccharide chain. This pocket may allow for accommodation of a second acetyl group where the xylose moiety is decorated at both C2 and C3. Although this may indicate that the esterase can utilise more bulky hydrophobic carboxylic acid groups than acetate, biochemical data indicate that the enzyme is specific for substrates comprising acetyl esters. However, due to the inability to obtain crystals of the esterase in complex with the appropriate oligosaccharide, a more detailed analysis of the structural features that confer substrate specificity is currently not possible.

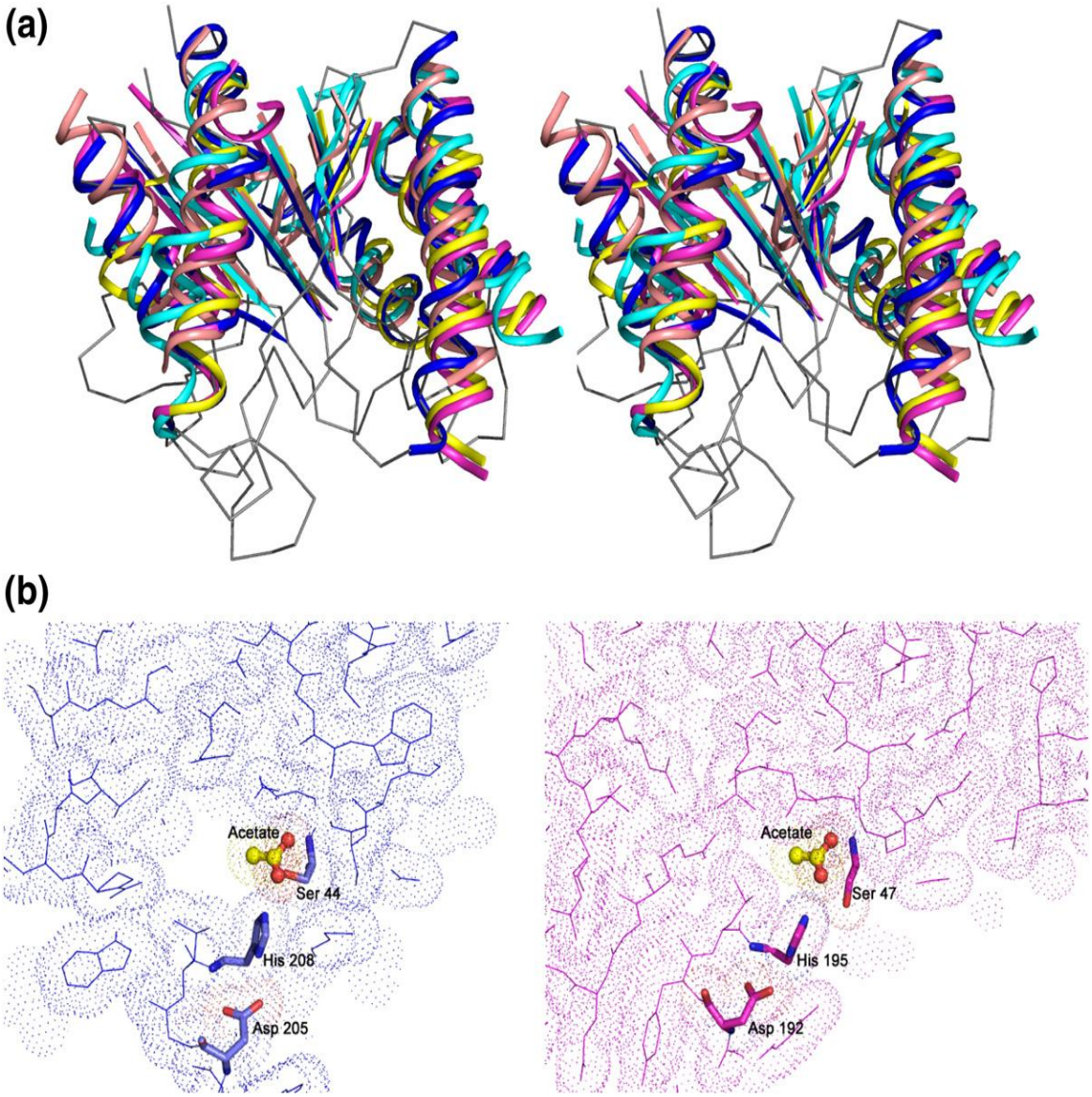


Figure 2.5 Comparison of the overall fold and active sites of enzymes displaying structural similarity to *Ct*Ces3-1.

(a) A stereo view of the superposition of five members of the α/β-hydrolase fold displaying closest structural similarity to *Ct*Ces3-1, displaying the core conserved secondary structural elements in cartoon format. Loops and non-conserved secondary structure have been omitted, with the exception of *Ct*Ces3-1, which is shown as a ribbon diagram in grey. Colours for the individual structures are *Ct*Ces3-1, dark blue; 1VJG (putative lipase from Nostoc sp.), pink; 1U8U (TAP L1 from Escherichia coli), cyan; 1WAB (platelet-activating acetylhydrolase from Bos taurus), yellow; 1YZF (lipase/ acylhydrolase from Enterococcus faecalis) salmon. (b) A comparison of the active site of *Ct*Ces3-1 (shown at the left in blue) with that of 1WAB (shown at the right in pink). The two enzymes are in the same orientation with dotted spheres representing the van der Waals' radii of the individual atoms, and are shown in cross-section (7 Å slab thickness) in an arbitrary plane that bisects the substrate-binding pocket. The acetate ion (shown throughout in ball-and-stick format) is modelled into its expected position in the active site of *Ct*Ces3-1, where it could make favourable interactions with the catalytic triad residues (shown throughout in stick format). The hydrophobic pocket that accommodates the methyl group is significantly larger in *Ct*Ces3-1 than in 1WAB.

A more general comparison of *Ct*Ces3-1 with the rest of the SGNH hydrolase superfamily reveals that the insertion of the 14 residue hairpin loop within β2 (residues 76–90) is a structural feature that is thus far unique to *Ct*Ces3-1. Interestingly, the two residues preceding this loop (Asp73 and Phe74) and the residues at positions 11 (Asp87) and 13 (Asp89) in this loop co-ordinate the bound calcium ion (Figure 2.4), indicating that metal ion binding may play an important role in the positioning of this loop and, as suggested previously, has a structural rather than a catalytic role in enzyme function. It is interesting to note, however, that the pattern of sequence conservation in close homologues of *Ct*Ces3-1 reveals that the three aspartate residues are highly conserved, suggesting that calcium ion binding may be a feature of other CE3s that are homologous to *Ct*Ces3-1. Thus, the crystal structure of *Ct*Ces3-1 indicates that the enzyme is a classic serine esterase containing a canonical catalytic triad in addition to two backbone nitrogen atoms and a side-chain amide that contribute to the stabilization of the tetrahedral transition state. The large internal cavity that extends from the active site may indicate how *Ct*Ces3-1 is able to accommodate more bulky polysaccharide substrates.

To investigate further the identity of the proposed catalytic triad, the *Ct*Ces3-1 mutants S44A, D205A, D205N and H208A were generated. The four mutants were expressed in *E. coli* and purified to electrophoretic homogeneity and their biochemical properties determined. The data (Table 2.3) show that the S44A and H208A mutants displayed no significant catalytic activity, supporting the proposed role for Ser44 as the catalytic nucleophile and His208 as the acid-base catalyst. Surprisingly, D205A and D205N were active against PNPAc, although the k_{cat}/K_m and K_m of the mutants were 7-20-fold and >10-fold lower, respectively, than the corresponding kinetic parameters of the wild type esterase (K_m of the wild type enzyme was >10 mM, see above). By contrast, D205A and D205N displayed no detectable activity against the more natural substrate, acetylated xylan. These data highlight the problems associated with utilizing non-natural subtrates with good leaving groups, such as 4-nitrophenolate, in determining the true roles of individual amino acids in the catalytic mechanism of hydrolytic enzymes. The decrease in K_m of the two aspartate mutants against PNPAc indicates that acylation (k_2) is much quicker than deacylation (k_3), suggesting a switch in the rate-limiting step of the reaction from k_2, in the wild-type esterase, to k_3 in the mutant. It is possible that in the aspartate mutants His208 displays reduced capacity to donate a proton (extracted from Ser44) to the leaving group during acylation. As the

leaving group in PNPAc does not require protonation, the mutation will have less effect on the hydrolysis of the aryl-acetate compared to acetylated xylan, where cleavage of the ester bond has an absolute requirement for protonation of the sugar leaving group.

Table 2.3 Catalytic activity of the wild type and active site mutants of *Ct*Ces3-1.

***Ct*Ces3-1**	**Substrate**	**K_{cat} (min^{-1})**	**K_m (mM)**	**K_{cat}/K_m ($M^{-1}min^{-1}x10^{-3}$)**
Wild Type	PNPAc	ND[a]	ND	152,7
S44A	PNPAc	NA[b]	NA	ND
D205A	PNPAc	39,9	1,8	22,1
D205N	PNPAc	10,5	1,5	7,2
H208A	PNPAc	NA	NA	NA
Wild Type	Acetylated xylan[c]	-	-	70,3
D205A	Acetylated xylan	-	-	NA
D205N	Acetylated xylan	-	-	NA

[a] ND, K_m and K_{cat} could not be determined individually as the catalytic rate and substrate concentration were directly proportional up to the maximum solubility of PNPAc (10mM)

[b] NA, no activity was detected

[c] Acetyl xylanase activity was determined at a substrate concentration of 10 mg ml^{-1} and the specific activity is reported as molecules of product produced per molecule of enzyme per minute.

2.4. CONCLUSIONS

This report demonstrates that *Ct*Ces3-1 is an acetyl esterase that displays a high level of specificity for acetylated xylan and xylooligosaccharides (Figure 2.6). This is entirely consistent with the other CE3 esterases characterized to date, which also hydrolyse acetylated xylan (Dalrymple *et al*, 1997; Aurilia *et al.*, 2000), indicating that the active site topographies of CE3 enzymes are likely to be highly conserved. The specificity of *Ct*Ces3-1 for xylan is consistent with the capacity of the cellulosome to degrade the plant cell wall, as this polysaccharide is the most abundant acetylated hemicellulose polymer in these composite structures (Bayer *et al.*, 2004). The likely biological function of *Ct*Ces3 is to increase access of the endo-acting xylanases for the undecorated xylan backbone. This may be particularly important as extensive deacetylation of xylans (that are not otherwise decorated), which is not associated with hydrolysis of the xylan backbone, will result in the precipitation of the polysaccharide. The location of xylanases and esterases within the cellulosome framework may ensure that deacetylation and glycosidic bond cleavage are tightly coupled and thus deacetylated xylooligosaccharides, which are highly soluble, will be generated rather than undecorated insoluble xylans. An intriguing feature of *Ct*Ces3 is that the enzyme contains two esterase modules that display 97% sequence identity and, by inference, will display the same substrate specificity. Indeed, the five differences in the two

modules, evident from the crystal structure of *Ct*Ces3-1 (Asn129/ Ser-345, Leu132/Val-348, Ile165/Val381, Lys215/ Thr430, Arg230/Lys446 in *Ct*Ces3-1 and *Ct*Ces3-2, respectively), are not located near the active site or the putative dimer interface but are mainly at, or close to, the surface of the esterase. If two catalytic modules can attack the same xylan chain, an avidity effect will occur (Freelove *et al.*, 2001; Boraston, *et al.*, 2002), and thus the enzyme will bind tightly to the polysaccharide. This cooperatvity is likely to confer specificity for xylan rather than xylooligosaccharides.

A

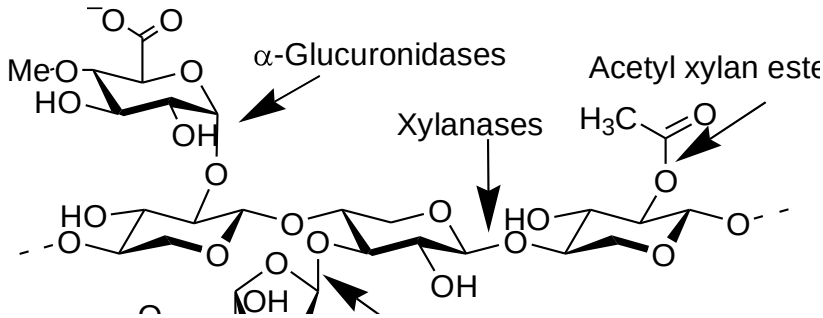


Figure 2.6 The enzymatic breakdown of xylan.
The structure of a decorated xylan molecule is presented, highlighting the enzymes and their recognized molecular bonds, which are cleaved during the complete hydrolysis of the complex polysaccharide.

CHAPTER 3 Evidence that family 35 carbohydrate-binding modules display conserved specificity but divergent function

Cedric Montanier[1], Alicia Lammerts van Bueren[2], Claire Dumon[1] , James E. Flint[1], Marcia A. Correia[3], Jose A. Prates[3], Susan J. Firbank[1], Richard[4] J. Lewis[1], Gilles G. Grondin[4], Mariana G. Ghinet[4], Tracey M. Gloster[5], Cecile Herve[6], J. Paul Knox[6], Brian G. Talbot[4], Johan P. Turkenburg[5], Janne Kerovuo[7], Ryszard Brzezinski[4], Carlos M. G. A. Fontes[3], Gideon J. Davies[5], Alisdair B. Boraston[2], and Harry J. Gilbert[1]

[1]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom; [2]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada V8W 3P6; [3]Interdisciplinary Centre of Research in Animal Health, Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Alto da Ajuda, 1300-477 Lisbon, Portugal; [4]Centre d'Étude et de Valorisation de la Diversité Microbienne, Sherbrooke, QC, Canada J1K 2R1; [5]Structural Biology Laboratory, The University of York, Heslington, York YO10 5YW, United Kingdom; [6]Centre for Plant Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom; and [7]Verenium Corporation, 4955 Directors Place, San Diego, CA 92121

Adapted from *Proceedings of the National Academy of Sciences* (2009) 106 (9), 3065–3070

Abstract

Enzymes that hydrolyse complex carbohydrates play an important role in numerous biological processes ranging from the maintenance of human health to the maintenance of marine and terrestrial life. These enzymes contain non-catalytic carbohydrate-binding modules (CBMs) that have an important substrate targeting function. In general there is tight correlation between the ligands recognized by the CBMs and the substrate specificity of the appended catalytic modules. Through high resolution structural studies, we demonstrate that the architecture of the ligand binding sites of four distinct CBMs, appended to three plant cell wall hydrolases and an enzyme, CsxA, that confers protection from, and metabolism of, an antibacterial fungal polysaccharide, is very highly conserved, which imparts similar binding specificity for glucuronic acid and/or Δ4,5-anhydrogalaturonic acid (Δ4,5-GalA). Δ4,5-GalA is released from pectin by the action of pectate lyases and as such acts as signature molecule for plant cell wall degradation. Thus, the CBMs appended to the three plant cell wall hydrolases, rather than targeting the substrates of the cognate catalytic modules, direct their appended enzymes to regions of the plant that is being actively degraded. Remarkably, the CBM, which is a component of CsxA, anchors the enzyme to the bacterial cell wall via its capacity to bind uronic acid sugars. This latter observation reveals a novel mechanism for bacterial cell-wall enzyme attachment. Overall, this work highlights how the biological role of CBMs is not dictated solely by their carbohydrate specificities but also by the context of their target ligands.

3.1. INTRODUTION

A generic feature of enzymes that hydrolyze complex carbohydrates is their modular structure. The most common noncatalytic modules are the carbohydrate-binding modules (CBMs), which are grouped into sequence-based families (Boraston *et al*., 2004). The general function of CBMs is to promote the interaction of the enzyme with the target substrate, thereby increasing the efficiency of catalysis (Boraston *et al*., 2003; Hall *et al.*, 1995; Tomme *et al.*¸1988). CBMs are prevalent in plant cell wall degrading enzymes where they direct the cognate catalytic modules to their target substrate within these complex composite structures (Boraston *et al*., 2004). The release of sugars from the recalcitrant carbohydrates of plant cell walls provides essential nutrients for microbial ecosystems that support marine and terrestrial life. Indeed, the ability of mammalian herbivores to access their major source of nutrients depends on the plant cell wall degrading activity of intestinal microorganisms (Hazlewood *et al.*, 1992). Within an industrial context the microbial enzymes that catalyze this process are integral to the exploitation of lignocellulose as an environmentally sustainable substrate for biofuel production (Ragauskas *et al.*, 2006). In addition to their contribution to plant cell wall degradation, there are increasing numbers of examples of CBMs playing a prominent role in other biological systems. For example, bacterial pathogens also deploy CBM-containing enzymes as virulence factors (Boraston *et al.*, 2006; Van Bueren *et al.*, 2007), whereas deficiencies in CBM-containing human proteins, such as phosphatases that regulate glycogen synthase (Cavanagh *et al.*, 1999; Wang *et al.*, 2002), can lead to serious diseases. CBMs are also located in bacterial enzymes that are components of protection systems. Thus, soil bacteria, such as *Amycolatopsis orientalis*, detoxify and metabolize the fungal polysaccharide chitosan by secreting an *endo*-chitosanase (Csn) and the CBM-containing *exo*-β-D-glucosaminidase, CsxA, which together depolymerize the polymer (Cote *et al.*, 2006; Sakai *et al.*, 1991).

As a general rule, the ligand specificity of bacterial CBMs reflects the substrates attacked by the cognate catalytic modules (Bollam *et al.*, 2004; Czjzek *et al.*, 2001; van Bueren *et al.*, 2005). An intriguing divergence from this general rule is provided by four family 35 CBMs (CBM35s) appended to CsxA and three different plant cell wall degrading enzymes (detailed in *Results*). These modules, despite being appended to enzymes with different substrate specificities, display extensive sequence similarity indicating that the CBMs recognize highly related ligands. Here, we show that all four CBM35s display specificity for Δ4,5-anhydrogalacturonic acid (Δ4,5-GalA), although

two of the proteins also interact with glucuronic acid (GlcA). X-ray crystallographic data reveal that the ligand binding site is highly conserved in the four CBM35s. Subtle modifications to the ligand binding site in two of the CBM35s enable these modules to discriminate between Δ4,5-GalA, a signature molecule for plant cell wall degradation, and GlcA, which is presented on the surface of several bacteria. By targeting Δ4,5-GalA, the CBM35s direct the plant cell wall hydrolases to accessible regions of the plant that are either undergoing biological degradation or remodeling. In contrast, the CBM35 appended to the defense enzyme CsxA appears to act as an adhesin molecule that anchors the glucosaminidase to the surface of the bacterium, rather than targeting the substrate chitosan. This report shows how CBMs that display conserved ligand specificities can exhibit divergent biological functions.

3.2. MATERIALS AND METHODS

3.2.1. Gene cloning and protein expression

The region of the genes encoding the CBM35 were amplified by PCR from genomic DNA or, in the case of Pel-CBM35 and Chi-CBM35, from the respective cloned genes, using the thermostable DNA polymerase Pfu Turbo (Stratagene) and primers listed in Table 3.1**.** Amplified DNA encoding Chi-CBM35 was cloned into *Nhe*I and *Xho*I restricted pET28a, whereas the other PCR products were cloned into *Nde*I and *Xho*I digested pET22b with exception of the Rhe-CBM35 that was cloned into *Nhe*I and *Xho*I restricted pET21a. All of the CBM35s contain a C-terminal His_6-tag except the Chi-CBM35.

Table 3.1 Primers used for the cloning of Rhe-CBM35.
Enzyme and restriction site are depicted in bold.

Plasmid	Protein		Primers (3' → 5')
pMC35A	Rhe-CBM35	F	CTC **GCT AGC** CCG GTA ATA TAT CAG GCT G
		R	CAC **CTC GAG** AGA TAC CGG CTG AAA GGC

3.2.2. Protein expression and purification of CBM35s

E. coli BL21 (STAR) DE3 cells harbouring the Xyl-CBM35, Pel-CBM35 and Rhe-CBM35 encoding recombinant expression vectors were cultured in Luria-Bertani broth at 37 °C to mid-exponential phase (A_{600nm} 0.6) and recombinant protein expression was induced by the addition of 1 mM isopropyl 1-thio-β-D-galactopyranoside and incubation for a further five hours at 37 °C. The CBM35s, derived from the plant cell

wall degrading enzymes, were purified by immobilized metal ion affinity chromatography (IMAC) using Talon^{TM} resin (Clontech) and elution in 20 mM Tris/HCl buffer containing 300 mM NaCl and 100 mM imidazole. An example of Rhe-CBM35 purification by affinity chromatography is presented in Figure 3.1. The eluted CBMs were then dialyzed against 20 mM Tris/HCl buffer pH 8.0 (Buffer A) and applied to a Bio-Rad Q12 column. The CBM35s were eluted with a 400 ml 0-500 mM NaCl gradient in Buffer A. The fractions containing the recombinant proteins were concentrated using a 10kDa MWCO Vivaspin 20 centrifugal concentrator and applied to a Superdex^{TM} 200 26/60 Hiload^{TM} column (GE Healthcare) equilibrated in 10 mM Tris-HCl buffer, pH 8.0, containing 150 mM NaCl. Protein was eluted at a flow rate of 1 ml/min. Both chromatography steps employed a Bio-Rad FPLC system. To purify Chi-CBM35 the cell pellet from lysed cells was in 100 ml 8 M urea and, after removal of any insoluble material, was dialyzed 6 times against 25 volumes of 20mM Tris/HCl buffer, pH 8.0, containing 0.5 M NaCl, 2 mM $CaCl_2$ and 2 mM $MgCl_2$ over a period of 3 days. The protein was purified by IMAC using 2 ml of Ni^{2+} affinity resin (GE Helthcare) and concentrated in a stirred cell ultrafiltration device with a 5000 MWCO membrane (Millipore). SDS/PAGE showed that the four CBM35s were >95 % pure.

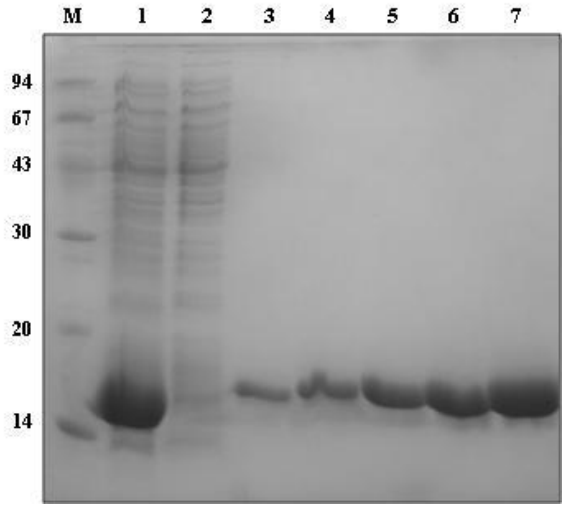


Figure 3.1 Purification of recombinant Rhe-CBM35 through affinity chromatography evaluated through SDS-PAGE in a 14% (w/v) polyacrylamide gel.

Lane M, Molecular-weight markers (kDa); Lane 1, cell-free extract after sonication; Lane 2, eluate from the column after sample loading; lane 3 to 7, fractions 1 to 5 of purified protein after Ni affinity column chromatography.

3.2.3. Mutagenesis

Site-directed mutagenesis was carried out using the PCR-based QuikChange site-directed mutagenesis kit (Stratagene) according to the manufacturer's instructions, using the plasmid for de encoding protein Pel-CBM35 as the template and primers pairs that are listed in

Table 3.2.

Table 3.2 Primers used to mutate Pel-CBM35

Residues	Primers (3' → 5')
K21A-F	CGTGATTGATAAC**GCG**CATGCTGGCTATACC
K21A-R	GGTATAGCCAGCATG**CGC**GTTATCAATCACG
K21N-F	CGTGATTGATAAC**AAT**CATGCTGGCTATACC
K21N-R	GGTATAGCCAGCATG**ATT**GTTATCAATCACG
H22A-F	GATTGATAACAAA**GCG**GCTGGCTATACC
H22A-R	GGTATAGCCAGC**CGC**TTTGTTATCAATC
Y25A-F	GATAACAAACATGCTGGC**GCG**ACCGGTACAGGATTTATCG
Y25A-R	CGATAAATCCTGTACCGGT**CGC**GCCAGCATGTTTGTTATC
R67A-F	GGCACAAGCGCA**GCG**CGTGCAACAGTAG
R67A-R	CTACTGTTGCACG**CGC**TGCGCTTGTGCC
W89A-F	CACCAACAGTAAT**GCG**ACCCAATGGCAAAC
W89A-R	GTTTGCCATTGGGT**CGC**ATTACTGTTGGTG
D116A-F	GCCGAGACTGCA**GCG**GGTTTAGCC
D116A-R	GGCTAAACC**CGC**TGCAGTCTCGGC
D116N-F	GCCGAGACTGCA**AAT**GGTTTAGCC
D116N-R	GGCTAAACC**ATT**TGCAGTCTCGGC
N120A-F	GATGGTTTAGCC**GCG**ATTGATAGTATTCGC
N120A-R	GCGAATACTATCAAT**CGC**GGCTAAACCATC

3.2.4. Carbohydrate and metal binding studies

The uronic acids used in the binding experiments were obtained as follows: GlcA, GalA, di- and tri-GalA were purchased from Sigma, while Δ4,5-GalAα1,4Gal and Δ4,5-GalAα1,4Galα1,4Gal were obtained by treating homopolygalacturonic acid (Sigma) with the pectate lyase *Cj*Pel10A to completion (monitored by no further increase at A_{257nm}). Xylooligosaccharides decorated with GlcA were obtained by digesting oat spelt xylan (Sigma) to completion, monitored by HPLC, with the xylanase *Cj*Xyn10B and the arabinofuranosidases *Cj*Abf51A and the GH43 enzyme ddAXH (Megazyme). The products were purified by size exclusion chromatography, as described by Proctor et al. (Proctor *et al.*, 2005), and then by ion exchange using Dowex-100 resin. The smallest charged oligosaccharide fraction (designated Xyl-GlcA) that bound to Xyl-CBM35 was retained for further use in ITC and crystallography experiments. The binding of CBM35s to carbohydrates was determined by ITC, which was carried out in 50 mM HEPES-Na buffer, pH 8.0, containing 2 mM $CaCl_2$ unless

otherwise specified, at 25 °C. The concentrations of the ligands in the syringe were ~3 mM and the CBM35s in the reaction cell were at 80-100 μM. The binding of the CBM35s to metal ions was determined as described previously (Bolam *et al.*, 2004). The determined K_a and ΔH values were used to calculate ΔS from the standard thermodynamic equation.

3.2.5. Amycolatopsis orientalis cell-binding studies.

Each reaction contained 40 mg of lyophilized *A. orientalis* cells in 1 ml 20 mM Tris/HCl buffer, pH 8.0, containing 1% BSA and 2 mM $CaCl_2$ and finally suspended in 1 ml of the same buffer. His-tagged AoCBM35 (100 μg) was added along with the His6-tag probe, which consisted of 10 μl of 0.5 mg ml^{-1} AlexaFluor 680–labeled streptavidin, previously complexed with biotin-NTA Ni^{2+}, and allowed to incubate in the dark with constant stirring for 1 h. Cells were recovered by centrifugation, washed 5 x 1ml with 20mM 20 mM Tris/HCl, pH 8.0, 0.1% Tween 20, 2 mM $CaCl_2$, then spotted onto a nitrocellulose membrane and visualized on a LICOR Odyssey system at 700 nm.

3.2.6. Immunofluorescence microscopy.

Bacteria. Cells fixed and made permeable on glass slides were incubated for 120 min at room temperature with primary chicken anti-CsxA antibodies and rabbit anti-Chi-CBM35 antibodies or with rabbit anti-Csn antibodies at a 1:1,200 dilution in PBS. After several washes in PBS, primary antibodies were detected by incubation with goat antirabbit Alexa Fluor 488 IgG and goat anti-chicken Alexa Fluor 555 IgG at a 1:200 dilution in PBSB for 90 min at room temperature. After several washes in PBSD and in water the DNA was stained for 10 min with Hoechst 33342 at a 1:5,000 dilution. After several washes in water, slides were mounted in Prolong Gold anti-fade reagent. Pictures were taken using an inverted Olympus IX 70 microscope with a Cool SNAP-Pro *cf*. monochrome camera and Image-Pro Plus software or with an Olympus Fluoview FV 300 confocal system.

Plant cell walls. For enzyme treatments before the immunohistochemistry, pectic homogalacturonan hydrolysis was carried out by incubating sections with 10 μg ml^{-1} of the *Cellvibrio japonicas* pectate lyase *Cj*PelA in 50 mM CAPS, 2 mM CaCl2 buffer pH 10, or with 40 μg ml^{-1} of the *Aspergillus niger endo*-polygalacturonase M2 (Megazyme) in phosphate–acetate buffer pH 5.0. Treatments were performed for 2 h at

roomtemperature. Due to the CBM35s sensitivity to calcium, a Tris-buffered saline (TBS) containing 5mM CaCl2 and 5% (wt/vol) BSA (BSA/TBS-Ca) was used for the incubation steps. The CBM35s were used at 100 µg ml^{-1}. The washing steps were performed with the TBS-Ca buffer. Samples were washed, mounted, and observed as described previously (McCartney *et al.*, 2004; McCartney *et al.*, 2006).

3.2.7. Crystallisation and structure solution

Chi-CBM35. Chi-CBM35 was treated with thrombin to remove the His6-tag and the resulting cleaved protein was buffer exchanged into 5 mM Tris/HCl buffer pH 8.0. Crystals were obtained at 18 °C by the vapor diffusion hanging drop method with 20 mg ml^{-1} Chi-CBM35 in a mother liquor containing 20% PEG4000, 10% isopropanol, 0.1 M sodium Hepes pH 7.4, at a 1:1 ratio of protein solution/mother liquor. A heavy atom soak was used for phasing, where crystals were soaked in 1 M NaI in mother liquor for 30 min. Crystals were frozen at 113 K directly in the cryostream after a short soak in mother liquor supplemented with 20% ethylene glycol. Data were collected with a Rigaku R-Axis 4++ area detector coupled to a MM-002 X-ray generator with Osmic Blue optics and an Oxford Cryostream 700. An atomic resolution data set for native Chi-CBM35 was collected on beamline X8C at the National Synchrotron Light Source (Brookhaven National Laboratories, Upton, NY). Complexes of Chi-CBM35 with GlcA and Δ4,5-GalAα1,4Gal (purified as described previously) (Abbott *et al.*, 2007) were obtained by soaking crystals of Chi-CBM35 in mother liquor supplemented with 10 mM sugar. These crystals were cryoprotected and data collected as described for the native crystals.

Xyl-CBM35, Rhe-CBM35, and Pel-CBM35. Proteins were crystallized by vapor diffusion at 20 °C. Pel-CBM35 was buffer exchanged into water in a 20 ml stirred ultra filtration device using a 10 kDa MWCO membrane and concentrated to 9.4 mg mL^{-1}. Crystals were grown from 0.25 M sodium phosphate/citrate buffer, pH 4.2, 0.2 M NaCl, and 20% PEG8000 (condition 26 of the Qiagen JCSG derivative screen) in 4 days, and frozen in the mother liquor containing 20% ethylene glycol as cryoprotectant. Data were collected on beamline ID14-4 at the European Synchrotron Radiation Facility (ESRF; Grenoble, France) to 1.7 Å. Xyl-CBM35 was crystallized in 20% (wt/vol) PEG3350, 0.4 M thiocyanate, pH 6.9 (condition 14 of the Qiagen JCSG derivative screen) with protein at 24 mg ml^{-1} in 2 mM $CaCl_2$. To obtain crystals of Xyl-CBM35 in

complex with Xyl-GlcA the ligand was included in the same crystallization conditions at 2 mg ml^{-1}. Crystals of the apo protein appeared after 21 days, whereas in the presence of Xyl-GlcA, crystals appeared after ≈100 days. The crystals were cryoprotected with the mother liquor with the addition of 30% (vol/vol) PEG400. Data for the native and complexed protein were collected on beamline IO3 at the Diamond Light Source (Oxfordshire, United Kingdom) to 1.9 Å and on a Rigaku 007 MicroMax copper rotating anode using an RAXIS IV++ imaging plate detector to 1.6 Å, respectively. Native crystals of Rhe-CBM35 were crystallized from 0.1 M Hepes sodium pH 7.5, 0.8 M sodium phosphate monobasic, 0.8 M potassium phosphate monobasic with protein at 20 mg ml^{-1} (Figure 3.2), and crystals of Rhe-CBM35 in complex with Δ4,5-GalAα1,4GalA were obtain in a 1.6 M sodium citrate solution with protein at 20 mg ml^{-1} supplemented with 1% (wt/vol) pectate lyase-digested polygalacturonic acid and 2 mM calcium chloride before mixing with mother liquor. The crystals were cryoprotected with mother liquor supplemented with 2 mM calcium chloride, 1% (wt/vol) digested polygalacturonic acid. Data for the native protein were collected on beamline ID29 at the ESRF to 1.5 Å, and the complex collected on a Rigaku 007 MicroMax copper rotating anode using an RAXIS IV++ imaging plate detector to 1.4 Å

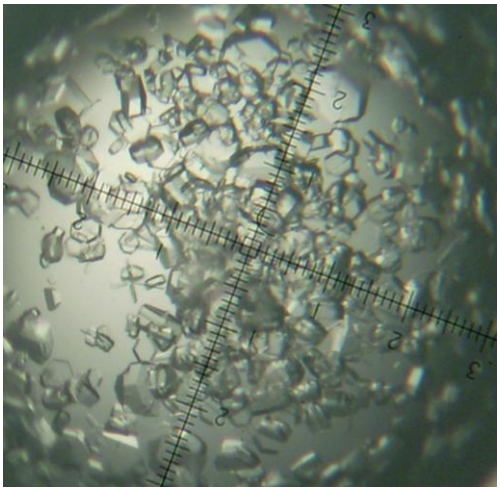


Figure 3.2 Crystals of Rhe-CBM35.
The Crystals were obtained by hanging-drop vapor diffusion in the presence of 0.1 M Hepes sodium pH 7.5, 0.8 M sodium phosphate monobasic, 0.8 M potassium phosphate monobasic

Data Processing and Structure Determination. Data for Chi-CBM35 were processed using Crystal Clear/d*trek (Pflugrath *et al.*, 1999) and all other computing was carried out with the CCP4 suite (Collaborative Computational Project Number 4, 1994) unless otherwise stated. ShelxD (Schneider *et al.*, 2002) was used to locate 5

iodide ions in the asymmetric unit using only the anomalous signal of the iodide derivative. Initial phasing to 3.0 Å was done by SAD using SHARP (Evans *et al.*, 2002). Density modification and phase extension was done with DM (Cowtan *et al.*, 1994). ARP/wARP (Murshudov *et al.*, 1997) was used for automatic model building followed by manual correction using COOT (Emsley *et al.*, 2004). Refinement was carried out with REFMAC (Lamzin *et al.*, 1997). The native model was used to solve the complex structures by molecular replacement using MOLREP (Vagin *et al.*, 2000). All data sets contained 2 molecules in the asymmetric unit. Water molecules were added using the REFMAC implementation of ARP/wARP and inspected visually before deposition. In all data sets, 5 percent of the observations were flagged as ‘‘free’’ (Brunger *et al.*, 1992) and used to monitor refinement procedures. Data for the native Rhe-CBM35 were integrated and scaled using DENZO and SCALEPACK (Otwinowski *et al.*, 1997) and for Pel-CBM35, Rhe-CBM35-Δ4,5-GalAα1,4Gal, Xyl-CBM35 (native and complexed with GlcA) using MOSFLM and SCALA (Evans *et al.*, 1993). The structures of Rhe-CBM35 and Xyl-CBM35 were determined by molecular replacement in MOLREP (Vagin *et al.*, 2000) and for Pel-CBM35 using AMoRe (Navaza *et al.*, 1994) using Chi-CBM35 as the search model. The starting models for Rhe-CBM35, Pel-CBM35, and Xyl-CBM35 were built automatically using ARP/wARP (Murshudov *et al.*, 1997), which was followed by rounds of manual rebuilding in COOT (Emsley *et al.*, 2004) interspersed with restrained refinement in REFMAC (Lamzin *et al.*, 1997). Solvent water molecules for Rhe-CBM35 and Pel-CBM35 were added using COOT, and checked manually. Xyl-CBM35–GlcA and Rhe-CBM35-Δ4,5-GalAα1,4GalA complexes were solved by molecular replacement using MOLREP with the coordinates for the native proteins used as search models and refined similarly to the native crystals. Water molecules were added using Arp_waters, and checked manually using COOT. Final model statistics for Rhe-CBM35 and Rhe-CBM35+Δ4,5-GalAα1,4Gal are given in (Table 3.3).Coordinates and structure factors have been deposited with the PDB codes of Pel-CBM35, 2W3J; Rhe-CBM35, 2W1W; Rhe-CBM35+Δ4,5-GalAα1,4Gal, 2W47; Xyl-CBM35, 2W46; Xyl-CBM35+GlcA, 2W87; Chi-CBM35, 2VZP; Chi-CBM35+GlcA, 2VZR; Chi-CBM35+Δ4,5-GalAα1,4Gal, 2VZQ for the 8 structures. All structural figures in the article were produced using Bobscript (Esnouf *et al.*, 1997).

Table 3.3 Data processing and refinement for Rhe-CBM35 and their complex

Data process	Rhe-CBM35 native	Rhe-CBM35+Δ4,5-GalAα1,4Gal
Data collection beamline, source	ESRF, ID-29	Rigaku 007, MicroMax copper rotating anode
Resolution: outer shell, Å	15.0–1.55 (1.61–1.55)	74.54–1.4 (1.43–1.40)
Space group	$P3_121$	$P6_522$
Unit cell parameters	a=b=46.4 Å, c=204.5 Å	a=b=86.0 Å, c=75.9 Å
R_{merge}, outer shell	0.10 (0.20)	0.08 (0.64)
Mean I/σI, outer shell	14.3 (9.0)	24.1 (3.7)
Completeness: outer shell, %	99.0 (100.0)	99.7 (98.3)
Multiplicity, outer shell	7.7 (7.2)	13 (12.5)
No. unique reflections	37854	31544
R_{cryst}	0.14	0.20
R_{free}	0.18	0.25
RMSD bonds, Å	0.016	0.007
RMSD angles, deg	1.58	1.19
PDB code	2W1W	2W47

3.3. RESULTS

3.3.1. Uronic Acid Recognition by CBM35.

CBMs generally display ligand specificities that reflect the substrates hydrolyzed by the appended catalytic modules (see Boraston *et al.*, 2004 for review). It is intriguing, therefore, that 4 CBM35s, which exhibit a high level of amino acid sequence identity (>38%), are components of enzymes where the cognate catalytic modules display highly divergent substrate specificities. The enzymes containing these CBMs are as follows: Rhamnogalauronan acetyl esterase Rgae12A (Cthe_3141) from *Clostridium thermocellum* (Figure 3.3) (containing two catalytic module of CE function, one X70 module of unknown funtion and our CBM from family 35), a PL10 pectate lyase (Pel10) from an environmental isolate (Solbak *et al.*, 2005), the *exo*-β-D-glucosaminidase CsxA from *A. orientalis* (Cote *et al.*, 2006), and the xylanase *Cj*Xyn10B from *Cellvibrio japonicus* (Kellett *et al.*, 1990). The 4 CBM35s are designated, henceforth, as Rhe-CBM35 (Rgae12A), Pel-CBM35 (Pel10), Chi-CBM35 (CsxA), and Xyl-CBM35 (*Cj*Xyn10B).

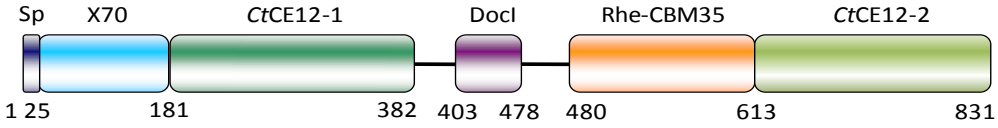


Figure 3.3 Molecular architecture of Rgae12A (Cthe_3141) of *Clostridium thermocellum*.
Domain bondaries are numbered with amino acid sequence. The abbreviated modules are as follows: Sp, signal peptide; DocI, type I dockerin modules.

Table 3.4 Thermodynamics of the binding of CBM35s to uronic acids

CBM35	Ligand	K_A(M^{-1})	ΔG (kcal $mole^{-1}$)	ΔH (kcal $mole^{-1}$)	$T\Delta S$ (kcal $mole^{-1}$)	*n*
Xyl-CBM35	GlcA	7.45×10^4	-6.68	-9.33	-2.68	1.0
Xyl-CBM35	Δ4,5-GalA[a]	2.33×10^5	-7.32	-5.16	2.16	1.2
Pel-CBM35	GlcA	NB[b]	-	-	-	-
Pel-CBM35	Δ4,5-GalA	2.00×10^4	-5.87	-6.95	-1.08	1.2
Pel-CBM35	OligoGalA[c]	NB	-	-	-	-
Rhe-CBM35	GlcA	2.28×10^3	-4.58	-6.34	-1.76	1.0
Rhe-CBM35	Δ4,5-GalA	6.14×10^4	-6.53	-7.20	-0.67	1.2
Chi-CBM35	GlcA	$1.89 (\pm 0.03) \times 10^4$	-5.83 (± 0.01)	-8.21 (± 0.10)	-2.37 (± 0.11)	1.16 ± 0.03

NB, no binding detected.

*Δ4,5-GalA is the digestion product produced by the action of *Cj*Pel10A on polygalacturonic acid which generates predominately the disaccharide Δ4,5-GalAα 1,4Gal and the trisaccharide Δ4,5-GalAα1,4GalΔ1,4Gal.

†The ligand comprised polygalacturonic acid hydrolyzed with polygalacturonase. The products are: GalA, GalA_ 1,4Gal, and GalAα 1,4Galα 1,4Gal.

Although a previous study showed that Xyl-CBM35 bound to an unknown component of xylans (Bolam *et al*., 2004), the specificities of the other CBM35s have not been investigated. Here, isothermal titration calorimetry (ITC) shows that all of the CBM35s appended to plant cell wall degrading enzymes bind to polygalacturonic acid predigested with a pectate lyase, but the proteins do not interact with the intact polysaccharide. The CBMs do not bind to pectin degraded with a hydrolytic polygalacturonase (Table 3.4), hinting that Δ4,5-GalA, the product of the β-elimination action (Figure 3.4), is the target ligand. Crystallographic data revealed that Chi-CBM35 also bound to the reaction products released from homogalacturonan by a pectate lyase. These data suggest that all 4 CBM35s recognize Δ4,5-GalA, as the anhydro sugar is located at the nonreducing end of the oligosaccharides generated by pectate lyase action (Moran *et al*., 1968). These CBM35s do not bind to GalA or a range of neutral sugars or their corresponding oligosaccharides (Table 3.6). ITC shows that Xyl-CBM35 and Chi-CBM35 also bind GlcA and Rhe-CBM35 interacts weakly with the uronic acid, whereas Pel-CBM35 displays no affinity for this sugar. Pel-CBM35 bound to calcium and, consistent with previous studies on Xyl-CBM35, all of the CBM35s derived from the plant cell wall degrading enzymes required the metal ion to bind their carbohydrate ligands (Table 3.5). GlcA and Δ4,5-GalA recognition by the CBMs is dominated by enthalpic forces with the change in entropy making an unfavorable contribution to ligand binding (Table 3.4). This thermodynamic signature is typical of the binding of proteins to soluble carbohydrates (reviewed in Boraston *et al.*, 2004).

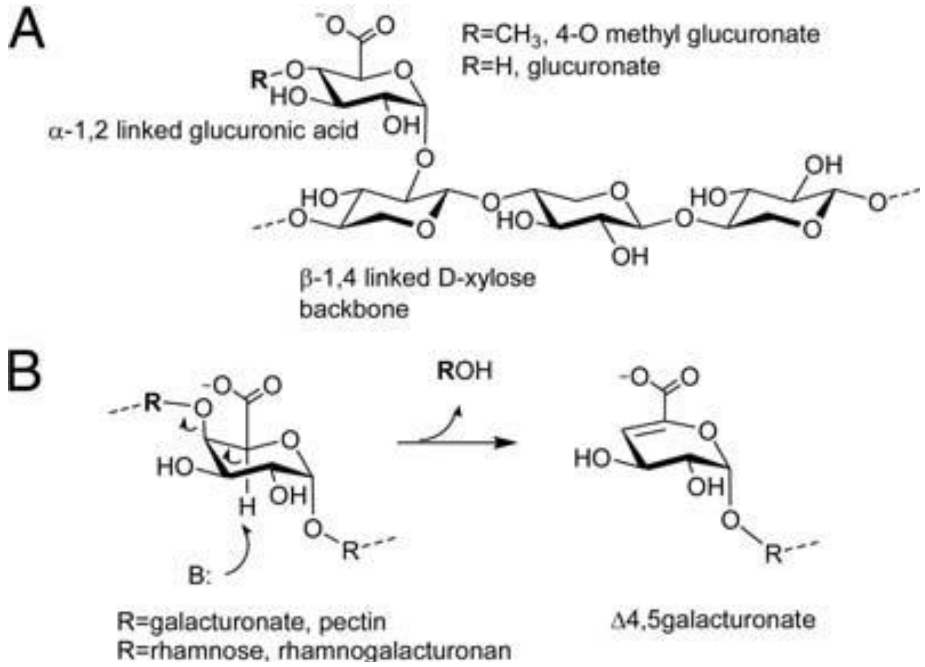


Figure 3.4 Ligands targeted by the CBM35 domains described in this work.
(*A*) α-1,2 linked GlcA moiety in glucuronoxylan is shown. (*B*) Δ4,5GalA (4,5anhydrogalactosyl) moiety produced by the action of lyases (β-eliminases) on pectin (a polysaccharide of α-1,4 linked galacturonides) or rhamnogalacturonan is revealed.

Table 3.5 Affinity of wild type and mutants of Pel-CBM35 for metal and carbohydrate ligands

Protein	Ligand	K_A (M^{-1})
Apo Pel-CBM35[a]	Calcium	3.21×10^4
Pel-CBM35 + EDTA[b] + Calcium[c]	Δ4,5-GalA	1.72×10^4
Xyl-CBM35 + EDTA[b] + Calcium[c]	Δ4,5-GalA	1.83×10^5
Rhe-CBM35 + EDTA[b] + Calcium[c]	Δ4,5-GalA	4.26×10^4
Pel-CBM35 D32A/K21N	Δ4,5-GalA	4.10×10^3
Pel-CBM35 D32A/K21N	GlcA	2.71×10^3
Pel-CBM35 D32N	GlcA	Weak binding ($<10^3$)
Pel-CBM35 D32N	Δ4,5-GalA	1.12×10^4
Pel-CBM35 K21N	GlcA	NB
Pel-CBM35 K21A	Δ4,5-GalA	2.62×10^4

EDTA was at 2 mM and calcium was at 10 mM. NB, no binding detected.
*Apo Pel-CBM35 was treated with Chelex-100 to remove any bound divalent metal ions.

Δ4,5-GalA, appearing as a result of pectate degradation or remodeling, is a signature molecule for plant cell wall damage. Thus, the likely function of the CBM35s located in the plant cell wall hydrolases is to target the cognate enzymes to regions of the plant that are particularly accessible to enzyme attack (see *Discussion*). It is also evident that Xyl-CBM35 binds to xylan but does not interact with unsubstituted

xylooligosaccharides (Bolam *et al*., 2004). To explore the unusual basis for xylan recognition, the polysaccharide was digested with xylanases and arabinofuranosidases to completion, and the products were fractionated (see Materials and Methods). ITC experiments identified a negatively charged oligosaccharide fraction containing GlcA (identified by HPLC analysis of the component sugars) bound to Xyl-CBM35. Thus, the capacity of Xyl-CBM35 to recognize xylan is likely to reflect the presence of GlcA side chains on the polymer. The protein, however, did not recognize 4-methyl-D-glucuronic acid (MeGlcA) (Table 3.6). Recent studies have shown that the ratio of GlcA and MeGlcA in xylans depends on the balance between the rate of glucuronoxylan synthesis and the rate at which the GlcA side chains in nascent glucuronoxylan are methylated (Pena *et al*., 2007). It has been proposed that in model plants, such as *Arabidopsis*, the rate of GlcA methylation is lower than glucuronoxylan synthesis, explaining the presence of significant quantities of the unmethylated uronic acid (Kauss *et al*., 1967). Although the biological rationale for targeting the unmethylated uronic is unclear, we speculate that regions of the glucuronoxylans containing GlcA rather than MeGlcA may have a more open structure. Thus, the Xyl-CBM35 may be directing its appended catalytic modules to regions of the hemicellulose that are particularly susceptible to enzyme degradation.

Table 3.6 Assessment of the binding of CBM35s to sugars

Protein	Sugar	*Binding*
Xyl-CBM35	4-Methyl-glucuonic acid	No binding detected
Pel-CBM35	4-Methyl-glucuonic acid	No binding detected
Rhe-CBM35	4-Methyl-glucuonic acid	No binding detected
Xyl-CBM35	Galacturonic acid	No binding detected
Pel-CBM35	Galacturonic acid	No binding detected
Rhe-CBM35	Galacturonic acid	No binding detected
Xyl-CBM35	Mannose	No binding detected
Xyl-CBM35	Mannohexaose	No binding detected
Xyl-CBM35	Galactomannan	No binding detected
Xyl-CBM35	Glucomannan	No binding detected
Xyl-CBM35	Glucose	No binding detected
Xyl-CBM35	Cellohexaose	No binding detected
Xyl-CBM35	Laminarin	No binding detected
Xyl-CBM35	Glucomannan	No binding detected
Xyl-CBM35	Galactose	No binding detected
Xyl-CBM35	Galactan	No binding detected

The binding of the CBM35s to the sugars listed was by isothermal titration calorimetry using the sugars or oligosaccharides at 10mMand the polysaccharides at 2mg/ml. The protein, which is in the cell, was at 50 µM. Titrations were carried out in the presence of 2 mM calcium in 50 mM sodium Hepes buffer, pH 7.5.

3.3.2. Chi-CBM35 Is a Bacterial Adhesion Molecule.

The data described above indicate that Chi-CBM35 does not target CsxA to its substrate chitosan, which lacks uronic acid. GlcA, however, is a component of some bacterial cell walls (Chatterjee *et al*., 1987, 1988) and thus Chi-CBM35 may anchor CsxA to the *A. orientalis* cell wall through an interaction with the exposed surface polysaccharides. This view is supported by in vivo localization studies that revealed both the catalytic module and CBM of CsxA are closely associated with the perimeter of *A. orientalis* cells (Figure 3.5 *Ai*). In contrast, Csn, a protein known to be secreted into the extracellular milieu (12), did not colocalize with *A. orientalis* cells, but rather was found in diffuse halos around the cells and associated with extracellular foci that were likely to be chitosan particles (Figure 3.5 *Aii*). Washing the *A. orientalis* cells with GlcA, but not glucose, removed endogenous CsxA from the cell surface of the bacterium (Figure 3.5 *Aiii* and *Aiv*).

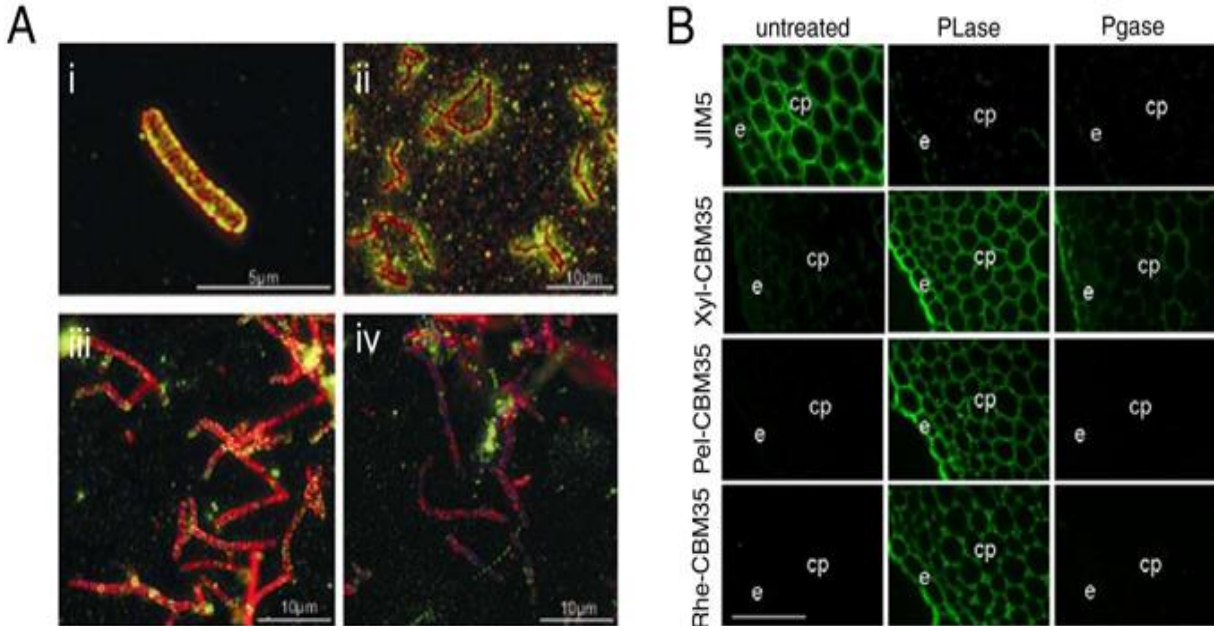


Figure 3.5 Binding of CBM35s to their target ligands.
(*AUpper*) Binding of Chi-CBM35 to the surface of *A. orientalis*. The confocal immunofluorescence image shows CsxA (red), Chi-CBM35 (green), and colocalization of the 2 proteins (orange) at the *A. orientalis* cell wall level (*i*). An epi-fluorescence image reveals the difference in cellular localization between a known secreted protein, Csn (green), and CsxA (red) (*ii*). **(*A Lower*)** Epi-fluorescence images of CsxA (red) and Chi-CBM35 (green) colocalization at the cell wall level. The bacterial DNA was stained in blue. Before heat fixation the mycelia of *A. orientalis* were washed twice for 10 min, with 100mM glucose (*iii*) or 100mM GlcA (*iv*). (Magnification: *i*, 5,000x; *ii*, 1,250x; *iii* and *iv*, 1,380x.) **(*B*)** Binding of CBM35s to cell walls of tobacco stems. Cell walls were either untreated or incubated with pectate lyase (PL) or polygalacturonase (PG) before incubation with the CBM35s. The regions corresponding toepidermis (e) and cortical parenchyma (cp) cell walls are indicated. (Scale bars, 10 µm)

CsxA is predicted to be a secreted protein with a Grampositive signal peptide cleavage site after amino acid 32 (Cote *et al*., 2006) but does not display any of the known Gram-positive cell wall immobilization motifs. Collectively, these observations provide compelling evidence that CsxA is a secreted protein that can be noncovalently attached to the cell wall of *A. orientalis* via a CBM-carbohydrate interaction. This cell wall association is most likely through an interaction of the CBM with GlcA, which is known to be a component of the glycolipids of other Actinomycetales (Chatterjee *et al*., 1987, 1988), although the biological ligand of the CBM may also be a related but unidentified uronic acid. By contrast immunohistochemical studies showed that none of the *C. japonicas* CBM35s bound to the cell wall of the host bacterium (Table 3.7).

Table 3.7 Evaluation of the binding of Xyl-CBM35 to Cellvibrio japonicas cells

Cellular fraction	Detection of Xyl-CBM35 added to a cell suspension of *Cellvibrio japonicus*
Cells recovered by centrifugation	No Xyl-CBM35 detected
Supernatant	Xyl-CBM35 detected

Cells of Xyl-CBM35 derived from a 100 mL overnight culture, after recovery by centrifugation, were resuspended in 1/10 volume of PBS and incubated with Xyl-CBM35 added to a final concentration of 10 μM for 2 h. The cells were removed by centrifugation and proteins in both the recovered cells and in the supernatant were subjected to Western analysis using an anti-His6 antibody, which recognizes the His-tagged Xyl-CBM35, as the probe.

3.3.3. CBM35 Targeting of Enzymes to Degraded Regions of the Plant Cell Wall.

By contrast to Chi-CBM35, the CBM35s appended to plant cell wall hydrolases are likely to play a role in substrate targeting. Indeed, the capacity of the CBM35s to bind to pectins, which have been depolymerized by pectate lyases, indicates that the role of the modules is to direct their cognate enzymes to those regions of the plant cell wall that are being actively degraded. To test this hypothesis, the capacity of the various CBM35s to bind to transverse sections of tobacco stems, treated with a pectate lyase and a polygalacturonase, was assessed. Xyl-CBM35, Pel-CBM35, and Rhe-CBM35 bind only very weakly or not at all to epidermal and cortical parenchyma cell walls of untreated sections. All 3 CBM35s, however, bind extensively to tobacco stem cell walls treated with pectate lyase, indicating that the proteins target the Δ4,5-GalA generated during pectin degradation (Figure 3.5 B). To ensure that the degradation of the pectins did not uncover ‘‘natural’’ ligands recognized by these proteins, plant cell walls were treated with a polygalacturonase. This enzyme hydrolyses the glycosidic bond in

polygalacturonic acid (homogalacturonan) pectins and thus the nonreducing sugar generated comprises GalA and not the anhydro sugar Δ4,5-GalA moiety. Degradation of pectins was demonstrated by the complete loss of binding of a homogalacturonan antibody to tobacco cells treated with either the pectate lyase or polygalacturonase (Figure 3.5 B). Pel-CBM35 and Rhe-CBM35, however, displayed no binding to the polygalacturonase-treated tobacco cell walls (Figure 3.5 B). Xyl-CBM35 did bind weakly to epidermal and cortical parenchyma cell walls after polygalacturonase treatment reflecting the exposure of xylan epitopes by pectic homogalacturonan removal (Figure 3.5 B). These data show that all 3 CBM35s target regions of the plant cell wall containing high concentrations of Δ4,5-GalA, although the Xyl-CBM35 can also direct enzymes to regions of these composite structures, presumably decorated xylans that contain GlcA.

3.3.4. Structure of Unliganded CBM35s.

The crystal structures of the 4 CBM35s showed that the proteins were structurally homologous (Figure 3.6). Indeed SSM (Secondary Structure Matching program; http://www.ebi.ac.uk/msd-srv/ssm/) analysis showed that Chi- CBM35 displays an rmsd of 1.1 Å with Pel-CBM35 over 121 C_α atoms, an rmsd of 0.8 Å with Rhe-CBM35 over 126 C_α atoms, and an rmsd of 1.1 Å with Xyl-CBM35 over 121 C_α atoms. In common with many CBM and lectin families, the 4 CBM35s display a jelly roll/β-sandwich fold comprising 2 antiparallel sheets consisting of 4 and 5 antiparallel β-strands, respectively. In all structures, except Rhe-CBM35, a metal ion is present, which has been modeled as calcium, based on its B-factor and coordination geometry exclusively with oxygen atoms. This calcium site is conserved in many lectin and several CBM families (reviewed in Boraston *et al.*, 2004). A second calcium is evident in all of the CBM35s, which is not structurally conserved in any other CBM family; the role of this metal in carbohydrate recognition is discussed below.

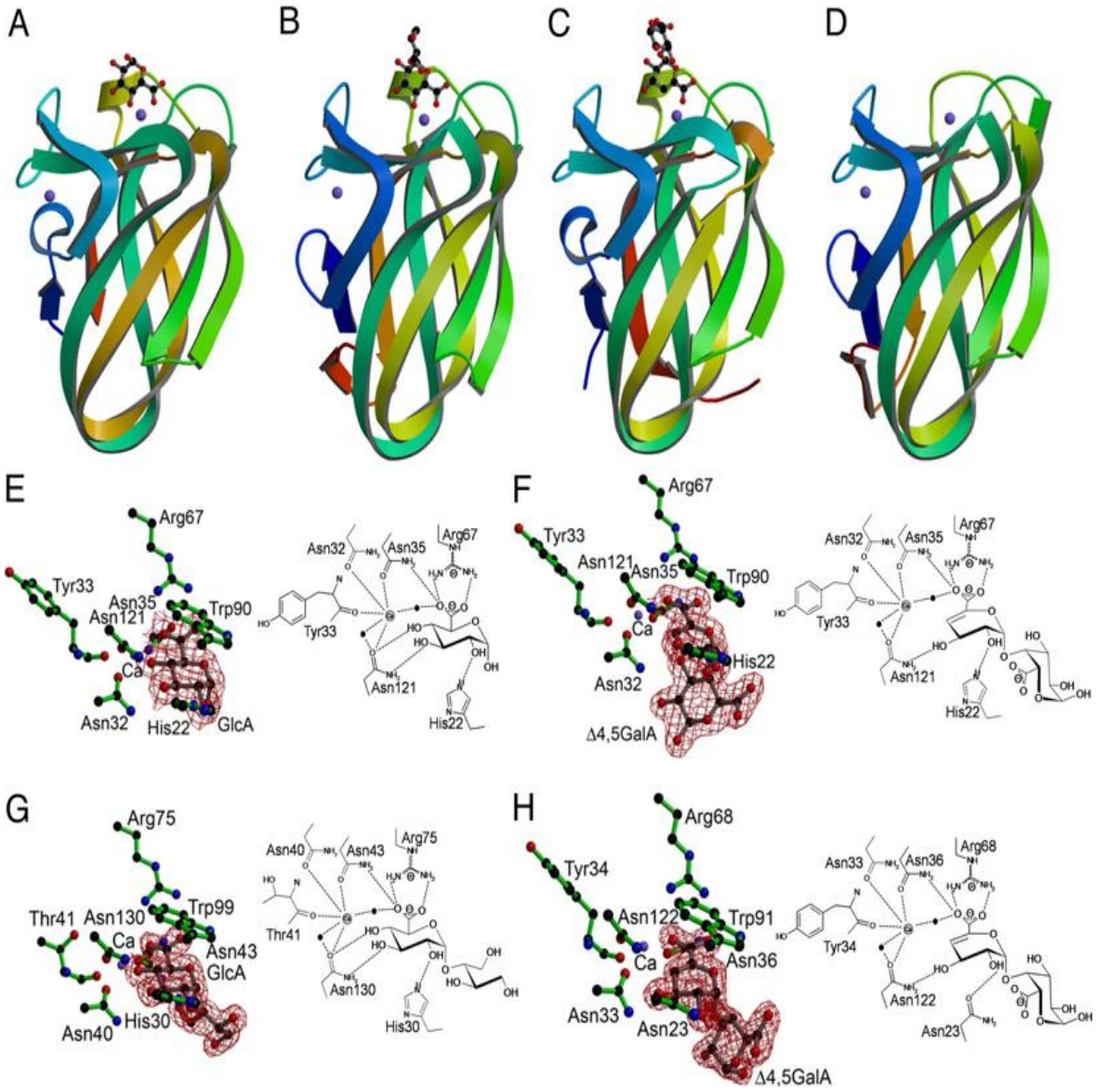


Figure 3.6 structure of family 35 CBMs.
(*A–D*) Ribbon representations, color-ramped from the *N*-terminus (blue) to *C*-terminus (red), of Chi-CBM35 (in complex with GlcA) (*A*), Xyl-CBM35 (in complex with a GlcA containing disaccharide) (*B*), Rhe-CBM35 (in complex with Δ4,5-GalAα1,4Gal) (*C*), and Pel-CBM35 (*D*). The calcium ions are represented as blue spheres and ligands in ball-and-stick representation. Structures are shown in identical orientations. (*E–H*) Binding sites of Chi-CBM35 in complex with GlcA (*E*), Chi-CBM35 in complex with Δ4,5-GalAα1,4Gal (*F*), Xyl-CBM35 in complex with a GlcA containing disaccharide (*G*), and Rhe-CBM35 in complex with Δ4,5-GalAα1,4Gal (*H*). (*E–H Left*) The structures of the active site with the observed electron density for the ligand are shown. The electron density maps are shown as maximum likelihood weighted $2F_{obs}$-F_{calc} maps contoured at 1σ. Hydrogen bonding schematics with the calcium atoms are shown as gray circles and water molecules coordinating the calcium are shown as black circles.

3.3.5. Structure of CBM35-Ligand Complexes.

The mechanism of ligand recognition was revealed from the crystal structures of Xyl-CBM35 in complex with a GlcA-containing decorated xylooligosaccharide, Rhe-CBM35 in complex with a Δ4,5-GalA-containing galacturonooligosaccharide, and Chi-

CBM35 bound to GlcA or Δ4,5-GalAα1,4Gal (Figure 3.6). In each structure the carbohydrate ligand is housed in a shallow surface indentation in which *O1* is solvent-exposed, which explains how the proteins are able to bind to the nonreducing end of pectin-derived oligosaccharides or the side chains of xylan. Thus, in Xyl-CBM35 the bound GlcA makes an α-linkage to another sugar that could not be identified with certainty from the electron density because of its disorder, whereas Δ4,5-GalA in complex with Rhe-CBM is α-1,4 linked to a D-GalA residue. The ligand binding site is highly conserved in the 4 CBM35s. Indeed, GlcA recognition by Xyl-CBM35 and Chi-CBM35 is identical (Figure 3.6 *E* and *G*), and thus the Xyl-CBM35-GlcA complex will be used as a general reference (see Figure 3.7 for an overlap of Xyl-CBM35 and Chi-CBM35 bound to uronic acid). The carboxylate at *C6* of GlcA is a key specificity determinant, explaining why these proteins recognize GlcA and not glucose, although productive polar interactions are also made with *O4*, *O3*, and *O2*. The pyranose ring also makes extensive hydrophobic interactions with an adjacent tryptophan. The *C6* carboxylate of the sugar makes bidentate hydrogen bonds with the 2 Nη atoms of Arg-75, whereas *O6A* also contacts a calcium ion that interacts with the protein through hydrogen bonds to the Oδ1 atoms of Asn-40 and Asn-43 and with the carbonyl of Thr-41. The inability of the R75A and N40A variants of Xyl-CBM35 and the equivalent mutants of Pel-CBM35 to bind either uronic acid (Table 3.5) demonstrates the importance of the arginine and calcium in carbohydrate recognition. *O4* makes hydrogen bonds with the Oδ1 atoms of Asn-40 and Asn-130 and also makes a polar contact with the calcium ion. *O3* interacts with Nδ2 of Asn-130 and *O2* contacts with the Nε2 of His-30 (Figure 3.6 G). The pyranose ring makes extensive hydrophobic contact with Trp-99. The importance of these interactions is consistent with the observation that the N130A, W99A, and H30A variants of Xyl-CBM35 and the W89A, H22A, N35A, and N120A mutants of Pel-CBM35 displayed no binding to either GlcA or Δ4,5-GalA (Table 3.6). The conservation of these amino acids in the 4 CBM35s provides further support for their pivotal role in carbohydrate recognition. The Xyl-CBM35 residues Asn-130, Trp-99, Arg-75, and Asn-43 are invariant in the 4 proteins; Asn-40 and His-30 are functionally conserved as Asn-40 is replaced by an aspartate in Pel-CBM35 and Asn-23 is structurally equivalent to His-30 in Rhe-CBM35. The weaker binding of Rhe-CBM35 to GlcA, compared with Xyl-CBM35 and Chi-CBM35, likely reflects the replacement of His-30 with Asn-23; the hydrogen bond between Asn-23 Oδ1 with *O2* is weaker than the polar contact between His-30 and *O2* (Figure 3.6 *G*

and *H*). Δ4,5-GalA binds at an identical site to GlcA in Chi-CBM35 (and by inference Xyl-CBM35 as its binding site is identical to Chi-CBM35; see Figure 3.7), and the interactions at *C2*, *C3*, *C5*, and *C6* are conserved (Figure 3.6 *E* and *F*). With Xyl-CBM35, the loss of the polar interaction at *C4* appears to be compensated by a tighter hydrophobic interaction between Trp-99 and the conjugated sugar ring in Δ4,5-GalA compared to the saturated pyranose ring in GlcA. Rhe-CBM35 displays a marked preference for Δ4,5-GalA, binding ≈30-fold more tightly than to GlcA. In contrast, Pel-CBM35 recognizes the pectin-derived ligand but displays no affinity for GlcA. The mechanism for this difference in specificity is subtle. Of the 4 CBMs, only Pel-CBM35 contains a lysine at position 21, which forms an ion pair with Asp-32. This interaction is predicted to prevent the aspartate from making a hydrogen bond with the critical *O4* moiety in GlcA. In the other 3 CBM35s the equivalent amino acids to Lys-21 and Asp-32 are both asparagine residues. Replacing Asp-32 and Lys-21 in Pel-CBM35 with asparagine residues restores GlcA recognition without compromising affinity for Δ4,5-GalA (Table 3.5).

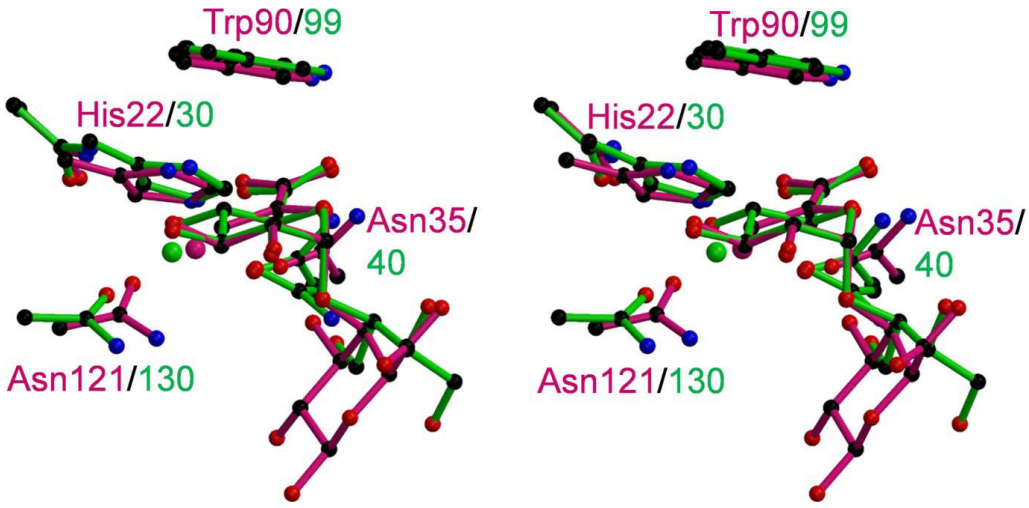


Figure 3.7 Comparison of the Chi-CBM35 and Xyl-CBM35 binding sites.
An overlay of the Chi-CBM35 binding site (pink) with Xyl-CBM35 (green) is shown in divergent stereo. Chi-CBM35 is shown in complex with Δ4,5-GalAα1,4Gal and Xyl-CBM35 in complex with a GlcA containing disaccharide. Calcium atoms are shown as spheres.

3.4. DISCUSSION

This report shows that despite displaying significant structural and biochemical similarities, there is considerable diversity in the biological functions of the 4 CBM35s. Although 3 of the CBM35s target the plant cell wall, notably Δ4,5-GalA, the capacity of Chi-CBM35 to function as a bacterial cell wall attachment module is perhaps the most

striking feature of this protein ensemble. Proteins can be retained on the surface of eukaryotic and prokaryotic cells through a variety of mechanisms that do not feature carbohydrate recognition. Thus, the observation that Chi-CBM35 tethers CsxA to the surface of *A. orientalis* by binding to a carbohydrate ligand is an unusual biological role for a CBM and carbohydrate binding proteins in general. It should be noted, however, that a recent report has shown that a family of carbohydrate-binding modules that bind to a wide spectrum of β-linked plant structural polysaccharides (CBM37), which are unique to *Ruminococcus albus*, anchor their cognate enzymes to the surface of the bacterium (Ezer *et al*., 2008). The target ligand for Chi-CBM35 in the cell wall of *A. orientalis* is most likely an uronic acid such as GlcA. Although the ITC-derived affinity of Chi- CBM35 for the uronic acid (KD ≈ 100 μM) is modest, this solution affinity does not reflect the binding strength on a membrane surface where clustering of the ligands can result in a marked increase in apparent affinity. The biological rationale for this function appears to be to keep the enzyme in proximity to the bacterium. By presenting the enzyme on the surface of the bacterium, only the cationic fragments of chitosan, which pose the most immediate threat to the bacterium, are degraded by CsxA. Furthermore, the monosaccharide products of hydrolysis would be in the direct vicinity of the bacterium promoting rapid transport into the bacterial cell for further metabolism. The cell adhesion role of CBMs, reported here and by Ezer *et al*. (Ezer *et al*., 2008), which was previously unconsidered, may prove to factor prominently in the function of these protein modules in the future. The capacity of the 3 CBM35s, which are components of plant cell wall degrading enzymes, to bind Δ4,5-GalA indicates that they play a role in targeting the hydrolases and esterases to regions of their composite substrate that are accessible to biological degradation. Pectins are arguably the most accessible region of the plant cell wall and indeed cloak other polysaccharides, such as xyloglucan, occluding them from enzyme attack (Marcus *et al*., 2008). Thus, by targeting enzymes to regions rich in Δ4,5-GalA, the CBM35s are directing their cognate catalytic modules to areas of the cell wall that are particularly accessible to further biological degradation. The origin of the pectate lyases that generate the Δ4,5-GalA molecules is unclear. It is possible that these enzymes are derived from saprophytic microorganisms that contribute to the plant cell wall degrading process. It is apparent, however, that plant genomes encode extensive families of pectate lyases (Henrissat *et al*., 2001), and thus it is possible that the anhydro sugar is released by endogenous enzymes. Although Δ4,5-GalA was not evident in the tobacco stem sections, suggesting

that pectate lyases were not expressed in these tissues, it is possible that these enzymes are only synthesized in specific cell types or differentiation states. It should be emphasized, however, that Xyl-CBM35 also displays significant affinity for GlcA, a side chain present in several xylans (Pena *et al.*, 2007), and thus this plasticity in uronic acid recognition provides the enzyme with a flexible targeting strategy that is tailored to the plant biomass presented to the bacterium. It is possible that Xyl-CBM35 initially directs the xylanase toward regions of cell walls that are being actively degraded but, as xylan structures are revealed, the enzyme is shuttled onto the hemicellulosic polysaccharide affording the enzyme access to its target substrate. It is interesting to note that the gene encoding the CBM35-containing xylanase is a component of a *C. japonicus* operon that also directs the synthesis of 2 other enzymes that contribute to xylan degradation; a GH62 arabinofuranosidase (Abf62A) and a CE1 ferulate/acetyl esterase (CE1A) (Kellett *et al*., 1990; Ferreira *et al*., 1993). Significantly, Abf62A and CE1 contain a CBM35 whose sequence is identical to Xyl-CBM35 and will therefore display specificity for Δ4,5-GalA and GlcA. Thus, the targeting of regions of plant cell walls that are being actively degraded is not restricted to a single xylanase, but is a feature of an enzyme ensemble that attacks decorated xylan structures. This report demonstrates how nature has exploited convergent binding specificity for divergent function. It is evident, therefore, that the context of the ligands recognized by CBMs is central to their biological function. In addition to providing insights into the biological role of CBMs as enzyme delivery molecules and host cell adhesins, respectively, this report also highlights the potential exploitation of these protein modules within an industrial setting. The novel specificities displayed by the family 35 CBMs will not only contribute to the toolbox of biocatalysts required to generate biofuels from lignocellulose, but also represent valuable probes that can be deployed in the analysis of plant cell wall architecture.

Acknowledgments:

Márcia Correia developed all the experiments that dealt with the biochemical and structural characterization of Rhe-CBM35. All remaining data presented in this chapter, particularly the characterization of additional CBM35 domains, were obtained by others.

CHAPTER 4 family 6 carbohydrate-binding modules display multiple β-1,3-linked glucan specific binding interfaces.

Márcia A.S. Correia[1], Virgínia M.R. Pires[1], Harry J. Gilbert[2,], David N. Bolam[2], Vânia O. Fernandes[1], Victor D. Alves[1], José A.M. Prates[1], Luís M.A. Ferreira[1] and Carlos M.G.A Fontes[1]

[1] *CIISA – Faculdade de Medicina Veterinária, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal;*[2] *Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom;*

Adapted from Federation of European Microbiological Societies (2009) 300(1), 48-57

Abstract

Non-catalytic carbohydrate-binding modules (CBMs), which are found in a variety of carbohydrate degrading enzymes, have been grouped into sequence-based families. CBMs, by recruiting their appended enzymes onto the surface of the target substrate, potentiate catalysis particularly against insoluble substrates. Family 6 CBMs (CBM6s) display unusual properties in that they present two potential ligand binding sites termed cleft A and B, respectively. Cleft B is located on the concave surface of the beta-sandwich fold while Cleft A, the more common binding site, is formed by the loops that connect the inner and the outer beta-sheets. Here we report the biochemical properties of CBM6-1 from *Cellvibrio mixtus* *Cm*Cel5A. The data reveal that CBM6-1 recognizes specifically β1,3-glucans through residues located both in cleft A and cleft B. In contrast, a previous report showed that a CBM6 derived from a *Bacillus halodurans* laminarinase binds to β1,3-glucans only in Cleft A. These studies reveal the different mechanisms by which a highly conserved protein platform can recognise a common ligand

.

4.1. INTRODUTION

Enzyme-substrate targeting is critical to the efficient hydrolysis of plant cell wall polysaccharides, a key process for carbon recycling within the biosphere. Plant cell walls are intricate macromolecules composed of a variety of polysaccharides that establish complex interactions, thus limiting substrate accessibility to enzymatic attack (Brett & Waldren, 1996). Plant cell wall hydrolases, therefore, play a crucial role in numerous biological processes, while also being of considerable industrial importance, particularly in the exploitation of lignocellulose as an environmentally sustainable substrate for biofuel production (Boudet *et al.,* 2003; Schell *et al.*, 2004; Ragauskas *et al.,* 2006). A generic feature of enzymes that hydrolyse complex carbohydrates is their modular structure; plant cell wall hydrolases contain catalytic and non-catalytic modules linked by flexible linker sequences. The most common non-catalytic modules found in these enzymes are Carbohydrate-Binding Modules (CBMs), which are grouped into sequence-based families (Coutinho & Henrissat., 1999). The general function of CBMs is to promote and maintain the interaction of the enzyme with the target substrate thereby increasing the efficiency of catalysis (Hall *et al.*, 1995; Bolam *et al.*, 1998; Fernandes *et al.*, 1999; Boraston *et al.*, 2003a). Based on the topology of the carbohydrate-binding site, CBMs have been classified into three types (Boraston *et al.*, 2004). The binding site of type A CBMs, which recognize crystalline polysaccharides, consists of three planar aromatic residues, a topology that is consistent with the planar conformation of the ligand. In contrast, type B and type C CBMs recognize single carbohydrate chains either internally or at the termini, respectively, and present a ligand specificity that reflects the substrate specificity of the appended catalytic domain (Boraston *et al.*, 2004). Structural studies revealed that type B and C CBMs accommodate their target ligands in clefts and pockets, respectively (Boraston *et al.*, 2004).

Family 6 CBMs (CBM6s) consist of a large family that display a variety of ligand specificities and adopt a canonical β-jelly roll fold. The surface of CBM6 proteins contain two potential ligand binding sites termed cleft A and cleft B (Czjzek e*t al.*, 2001; Boraston *et al.*, 2003c; Pires *et al.*, 2004; Henshaw *et al.*, 2006). One of these sites, cleft B, is located on the concave surface, while the second site, cleft A, is on the edge of the protein and is formed by the loops that connect the inner and the outer β-sheets. Although cleft B displays a topology that is typical of type B binding-sites, cleft A generally, but not exclusively, resembles a type C carbohydrate-binding region,

which recognises single sugars or the termini of polymeric carbohydrates (Van Bueren *et al.*, 2005). The variation in ligand specificity in CBM6 reflects, in part, the functionality of these two binding sites. Thus, CBM6s that bind xylan (Czjzek e*t al.*, 2001; Boraston *et al.*, 2003c), laminarin (Van Bueren *et al.*, 2005) and the non-reducing glucose of β-glucans (Pires *et al.*,2004; Henshaw *et al.*, 2006) have been described, with ligand binding occurring in cleft A. Although cleft A forms a deep cleft in CBM6s that bind to the internal regions of xylan, in the majority of proteins in this family, cleft A forms a pocket. This pocket-like topology is exemplified by a CBM6 from a *Bacillus halodurans* laminarinase; the protein module binds to the terminal sugar of the helical glucan in cleft A but the ligand extends across the surface of the protein (Van Bueren *et al.*, 2005). The only report of a functional cleft B binding site is in CBM6-2 from *Cm*Cel5B (Pires *et al.*, 2004; Henshaw *et al.*, 2006), where the concave surface recognizes the internal regions of β1,4- or mixed linked β1,3-β1,4-glucans, while cleft A binds to the non-reducing end of β-linked gluco and xylo-configured polymers.

Cellvibrio mixtus is a saprophytic mesophilic bacterium that colonizes the soil and efficiently degrades a variety of plant cell wall polysaccharides. An enzyme derived from this bacterium, *Cm*Cel5B, presents a modular arquitecture containing an N-terminal family 5 glycoside hydrolase (GH) catalytic module followed by an internal CBM6, termed CBM6-1, and a C-terminal CBM6, termed CBM6-2 (Fontes *et al.*, 1998). The structure of CBM6-2 revealed the presence of two functional ligand binding sites, cleft A and cleft B, respectively (Pires *et al.*, 2004; Henshaw *et al.*, 2006). Here we present the biochemical properties of *C. mixtus* CBM6-1. The data revealed that CBM6-1 binds exclusively to β1,3-glucans through residues located both at clefts A and B. Therefore, β1,3-glucans can interact with CBM6 modules at different binding sites highlighting the plasticity in carbohydrate recognition displayed by this protein fold.

4.2. MATERIALS AND METHODS

4.2.1. Protein expression and purification

Genes encoding the truncated derivatives of *Cm*Cel5B, CBM6-1, CBM6-2 and CBM6-1/2 were amplified by PCR from *C. mixtus* genomic DNA using the primers listed in Table 4.1 employing the thermostable DNA polymerase NZYPremium (NZYTech Ltd). Forward and reverse primers incorporated, respectively, *Eco*RI and *Xho*I restriction sites, which were used for the subsequent cloning into pET32a (Table

4.1). Amplified DNA was directly cloned into pNZY28 (NZYTech Ltd) and sequenced to ensure that no mutations were accumulated during the amplification. The genes were subsequently sub-cloned into *Eco*RI/*Xho*I restricted pET32a (Novagen) generating pCBM6-1, pCBM6-2, and pCBM6-1/2, respectively (Sambrook *et al.*, 1989). All recombinant derivatives contained an internal His_6-tag and consisted of recombinant thioredoxin fusion proteins. To engineer a gene encoding two copies of CBM6-2 organized in tandem, DNA encoding CBM6-2 and the linker sequence separating CBM6-1 and CBM6-2, was amplified from *C. mixtus* genomic DNA with the primers listed in Table 4.1. The forward primer contained an engineered *Sal*I restriction site. The resulting PCR fragment was subcloned from pNZY28 (NZYTech Ltd), on a *Sal*I-*Xho*I restriction fragment, into the *Xho*I site of pCBM6-2. The stop codon that separated the two genes was subsequently removed using the mutagenesis primers listed in (Table 4.1), allowing the production of a thioredoxin fusion protein containing two CBM6-2 copies organized in tandem. Genes encoding *Cm*Cel5B catalytic domain, GH5, and the full length enzyme, GH5-CBM6-1/2, were amplified from *C. mixtus* genomic DNA as described above and using the primers listed in (Table 4.1). Resulting PCR products were cloned into pNZY28 (NZYTech Ltd) and sequenced to ensure that no mutations had occurred during the amplification. The obtained recombinant plasmids were digested with *Nhe*I and *Xho*I and the excised DNA fragment was cloned into the similarly restricted expression vector pET21a, to generate pGH5 and pGH5-CBM6-1/2. The two recombinant proteins contained a C-terminal His6 tag. The molecular architectures of all the proteins generated in this study are presented in Figure 4.1.

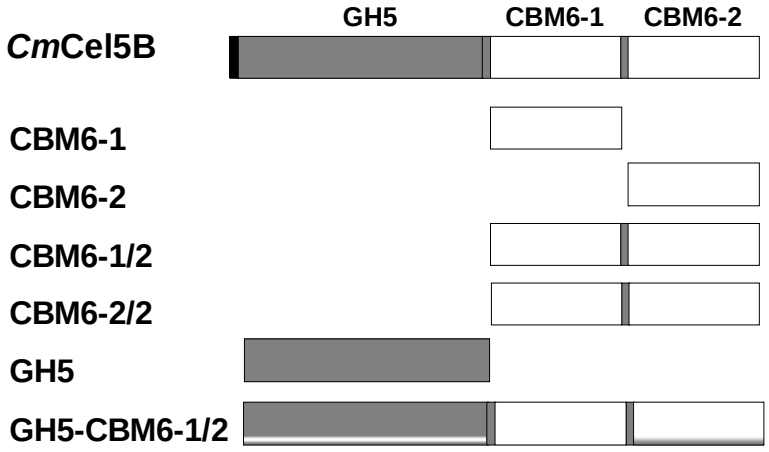


Figure 4.1 Molecular architecture of truncated derivatives of the *Cm*Cel5B used in this study.

Table 4.1 Primers used to obtain the genes encoding CmCel5B derivatives used in this work and for the mutagenesis of CBM6-1.
Engineered restriction sites and mutation points are depicted in bold.

Clone	*Sequence (5'→3')*	*Direction*
CBM6-1	CTC **GAA TTC** ACC TGC ACC AAA GCC AAT	FOR
	CAC **CTC GAG** TTA TTC TAC TTT CAG CCA GTT	REV
CBM6-2	CTC **GAA TTC** GTA ATC GCG ACT ATT CAG	FOR
	CAC **CTC GAG** TTA ATG TGT CTT GTT G	REV
CBM6-1/2	CTC **GAA TTC** ACC TGC ACC AAA GCC AAT	FOR
	CAC **CTC GAG** TTA ATG TGT CTT GTT G	REV
CBM6-2/2	CTC **GTC GAC** AGC ACT GGC GGT GAC AAC	FOR
	CAC **CTC GAG** TTA ATG TGT CTT GTT G	REV
CBM6-2/2-FL	CAT CAA CAA GAC ACA T**G**A AAG CAC TGG CGG TGA C	FOR
	GTC ACC GCC AGT GCT TT**C** ATG TGT CTT GTT GAT G	REV
GH5	CTC **GCT AGC** GTA CCC GCG TTA CAG GTG	FOR
	CAC **CTC GAG** ATC GAT GGG TTT ACG	REV
GH5-CBM6-1/2	CTC **GCT AGC** GTA CCC GCG TTA CAG GTG	FOR
	CAC **CTC GAG** ATG TGT CTT GTT GAT GAG	REV
CBM6-1 Y33A	GGT TTG AAT GTC GGC **GCT** ATC GAT GGC GGC GAC	FOR
	GTC GCC GCC ATC GAT **AGC** GCC GAT ATT CAA ACC	REV
CBM6-1 W92A	CCC GCG ACG GGC GGC **GCG** CAA AAC TGG CAG ACC	FOR
	GGT CTG CCA GTT TTG **CGC** GCC GCC CGT CGC GGG	REV
CBM6-1 E73A	GGA CAG CTC CAA TTG **GCA** AAA GCC GGT GGC AGC	FOR
	CT GCC ACC GGC TTT **TGC** CAA TTG GAG CTG TCC	REV
CBM6-2 E73A	CAG CTT GAC ATT T**GC A**GA AGC AGG CGG	FOR
	CCG CCT GCT TC**T GC**A AAT GTC AGG CTG	REV

E. coli Origami DE3 cells harbouring all the recombinant expression vectors, except pGH5 and pGH5-CBM6-1/2, were cultured in Luria-Bertani broth at 37 °C to mid-exponential phase (A_{600nm} 0.6) and recombinant protein expression was induced by the addition of 1 mM isopropyl 1-thio-β-D-galactopyranoside and incubation for a further sixteen hours at 19 °C. BL21 DE3 *E. coli* were used to express the catalytically active *Cm*Cel5B derivatives GH5 and GH5-CBM6-1/2, encoded by pGH5 and pGH5-CBM6-1/2, respectively, following the same procedure described above. Soluble recombinant proteins were purified by immobilized metal ion affinity chromatography as described previously (Carvalho *et al.*, 2003; Charnock *et al.*, 2000; Dias *et al.*, 2004) and buffer exchanged into 20 mM Tris-HCl buffer, pH 7.5, containing 100 mM NaCl and 5 mM $CaCl_2$ (Buffer A). SDS/PAGE showed that all the recombinant proteins were more than 95 % pure.

4.2.2. Source of sugars used

All soluble polysaccharides were purchased from Megazyme International (Bray, County Wicklow, Ireland), except oat spelt xylan, laminarin and hydroxyethylcellulose (HEC), which were obtained from Sigma. Avicel (PH101) was

obtained from Serva while acid-swollen cellulose was prepared as described previously (Najmudin *et al.*, 2006).

4.2.3. Mutagenesis

Site-directed mutagenesis was carried out using the PCR-based NZYMutagenesis site-directed mutagenesis kit (NZYTech Ltd) according to the manufacturer's instructions, using plasmids pCBM6-1 and pCBM6-1/2 as templates. The sequences of the primers used to generate these mutants are displayed in Table 4.1. The mutated DNA sequences were sequenced to ensure that only the appropriate mutations had been incorporated into the nucleic acid.

4.2.4. Affinity Gel Electrophoresis

The affinity of CBM6-1 and other *Cm*Cel5B derivatives for a range of soluble polysaccharides was determined by affinity gel electrophoresis (AGE). The method used was essentially that described by Tomme *et al.* (2000) using the polysaccharide ligands at a concentration of 0.1 % (w/v) unless stated otherwise. Electrophoresis was carried out for 4 h at room temperature in native polyacrylamide gels containing 10 % (w/v) acrylamide. The non-binding negative control protein was bovine serum albumin (BSA). Quantitative assessment of binding was carried out as described previously (Takeo, 1984), using polysaccharide concentrations ranging from 0.001 to 0.5 % (w/v).

4.2.5. Isothermal Titration Calorimetry

The binding of the CBMs to carbohydrates was quantified through Isothermal Titration Calorimetry (ITC). ITC was carried essentially as described previously (Flint *et al.*, 2004), except that proteins were dialyzed into 50 mM NaHepes, pH 7.5, containing 2 mM $CaCl_2$, unless otherwise specified, at 25 °C. The concentration of the ligands in the syringe was ~3 mM for the oligosaccharides and 0,5% (w/v) for the polysaccharides and the CBMs in the reaction cell were at 80-100 µM. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single site-binding model (Microcal ORIGIN, Ver. 5.0, Microcal Software, Northampton, MA). The molar concentration of CBM binding sites present in the polysaccharide ligands was determined as described previously (Bolam *et al.*, 2004). The fitted data yields the association constant (K_a) and the enthalpy of binding (ΔH).

Other thermodynamic parameters were calculated using the standard thermodynamic equation:

$$-RT\ln K_a = \Delta G = \Delta H - T\Delta S$$

All data show the average and standard deviation of three independent titrations.

4.2.6. Enzyme assays

Enzyme activity was determined at 37ºC in 20 mM Tris-HCl, pH 7,5, 20 mM NaCl, 5 mM $CaCl_2$ (Buffer A), using various polysaccharides as substrates and the rate of reducing sugar released was measured with the method described by Miller (Miller, 1959). The assay mixture (600 µl) contained 20 µl of enzyme (with appropriate dilution) and concentrations of substrate ranging from 0.15 to 2.00 % (w/v). The mixture was incubated at 37 ºC for 5, 10 and 15 min to confirm linearity of the reaction. All reported results are the mean of three separate experiments.

4.3. RESULTS AND DISCUSSION

4.3.1. Ligand specificity of CBM6-1

*Cm*Cel5B is a modular cellulase containing an N-terminal family 5 glycoside hydrolase (GH) catalytic domain followed by CBM6-1 and CBM6-2 (Van Bueren *et al.*, 2005). The structure of CBM6-2 revealed the presence of two ligand binding sites in the beta-sandwich structure, which were termed cleft A and cleft B (Pires *et al.*, 2004; Henshaw *et al.*, 2006). Cleft A was shown to bind to terminal glucose and xylose residues while cleft B recognizes cellulose and beta-1,4-1,3-mixed linked glucans (Henshaw *et al.*, 2006). The ligand specificity of CmCel5B CBM6-1 remains unknown and it is possible that both CBM6-1 and CBM6-2 act in concert during carbohydrate recognition. Alignment of CBM6-1, CBM6-2 and the laminarin-specific CBM6 of *B. halodurans*, suggests that CBM6-1 cleft A is indeed functional, since all residues involved in ligand recognition in CBM6-2 are conserved in CBM6-1 (Figure 4.2). In contrast, there are significant differences in the residues presented on the surface of cleft B in the two *Cm*Cel5B CBM6s. Significantly, major changes are observed in subsite 1 (CBM6-2 cleft B contains 4 sugar binding subsites) where Gly44, present in CBM6-2, is absent in CBM6-1 and CBM6-2 Gln110 is replaced by leucine in CBM6-1. More significantly, in subsite 4, Lys114, which makes several direct hydrogen bonds with the

sugar, is replaced by a serine in CBM6-1. These amino acid differences suggest that CBM6-1 cleft B should display a much restricted binding interface.

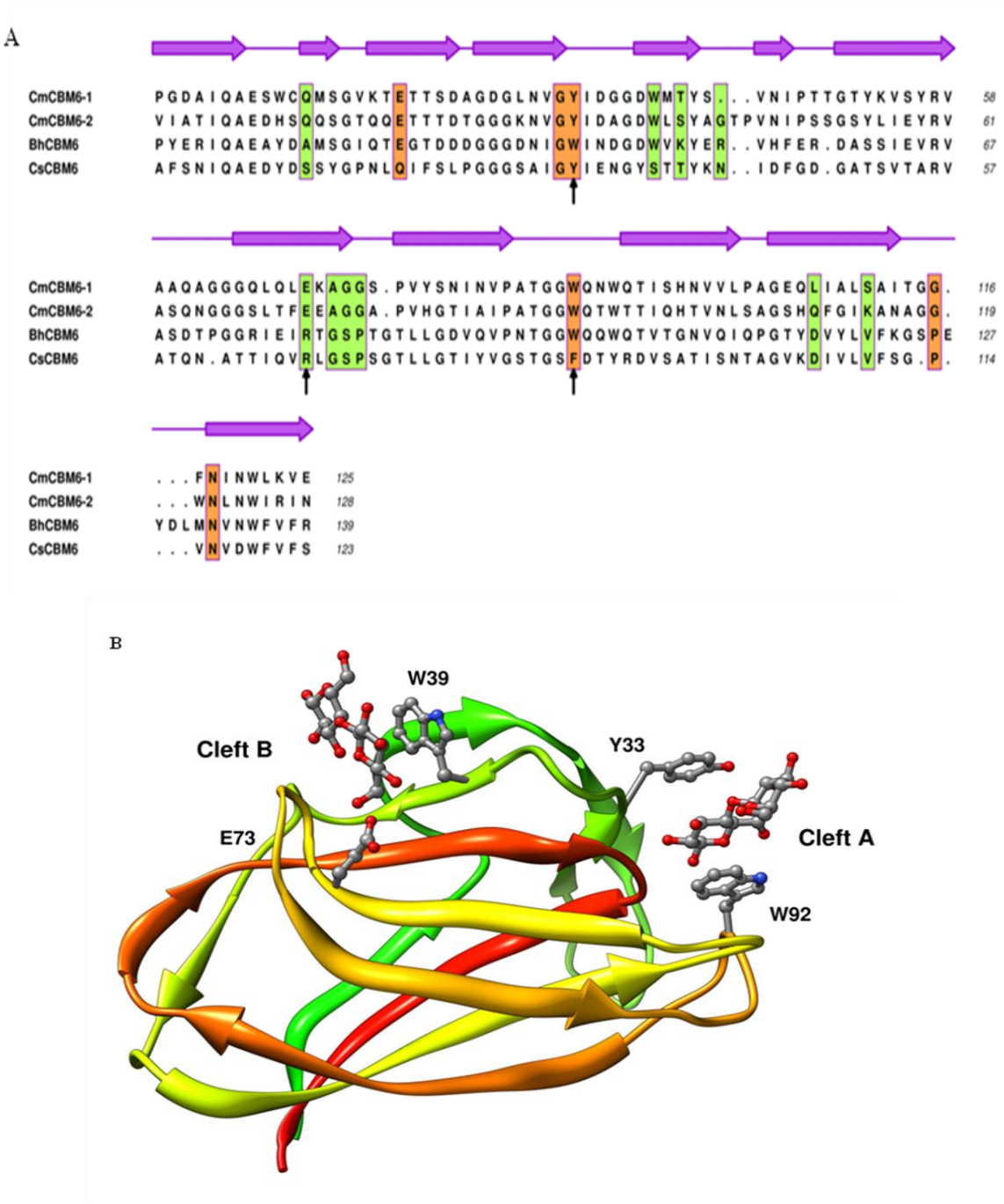
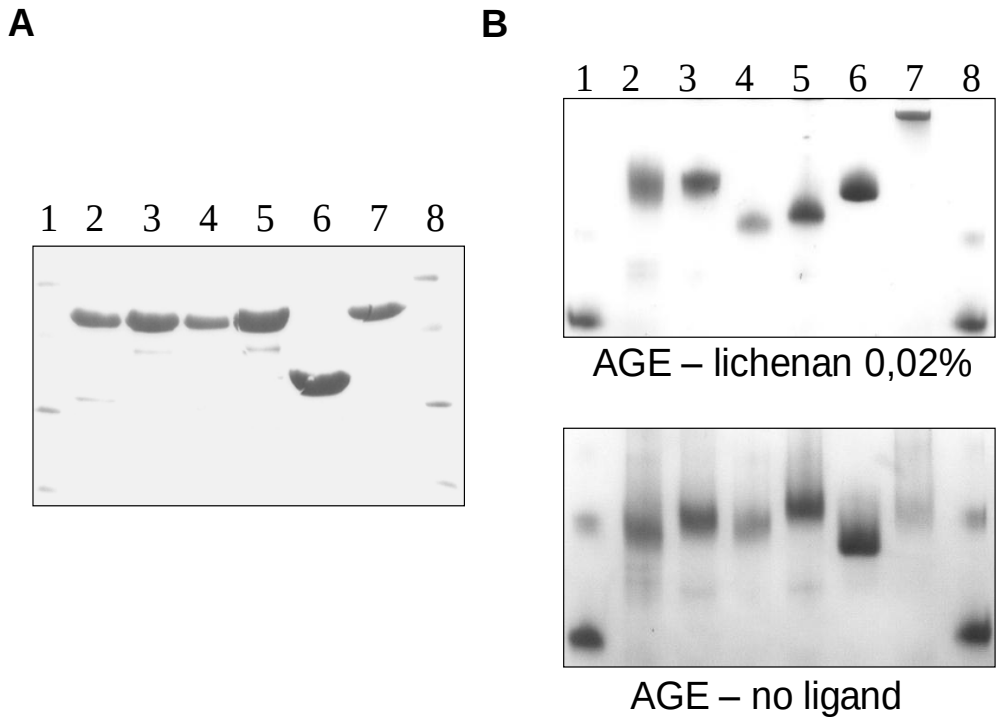


Figure 4.2. Structural alignment of CBM6-1 and CBM6-2 and Three-dimensional structure of *Cm*CBM6-2.

(A) Structural alignment of CBM6-1 and CBM6-2 with the laminarin specific CBM6 from *Bacillus halodurans* (*Bh*CBM6) and the xylan specific CBM6 from *Clostridium stercorarium* (*Cs*CBM6). Residues belonging to cleft A are boxed in orange and to cleft B in green. Arrows above the alignment represent the secondary structure elements of *C. mixtus* CBM6-2. Residues that were subjected to mutagenesis are highlighted with an arrow below the sequence. The protein alignment was prepared with CLUSTAL W (Thompson *et al.* 1994). **(B)** Three-dimensional structure of *Cellvibrio mixtus* CBM6-2 shown in color ramped format in complex with two cellobiose molecules bound to clefts A and B,

respectively (PDB ID: 1UYX). Amino acids dominating carbohydrate recognition at the two binding sites are indicated.

In order to clarify the functional implications of these amino acid substitutions in the ligand affinity of CBM6-1, the *Cm*Cel5B non-catalytic domain was purified to electrophoretic homogeneity and its ligand specificity was evaluated by ITC and AGE. The data, presented in Figure 4.3 and Figure 4.4, revealed that CBM6-1 binds to laminarin, with an affinity constant K_a of 4×10^3 (M^{-1}). Thus, CBM6-1 displays a relatively modest affinity for laminarin and determination of the thermodynamic parameters was not able to be performed with accuracy. However, in general, binding of CBMs to soluble saccharides is enthalpy-driven with entropy making an unfavourable contribution to ligand binding (Notenboom *et al.*, 2001; Boraston *et al.*, 2003b; Bueren *et al.*, 2005). Significantly AGE did not reveal any binding of CBM6-1 to laminarin, which may reflect the low degree of polymerization of the polysaccharide (Nisizawa *et al.*, 1963), and the stoichiometric binding of the polysaccharide to the ligand, which, collectively, would result in little electrophoretic retardation of the protein. A ligand specificity screen using AGE revealed that CBM6-1 displays no significant affinity for xyloglucan, the β1,4-β1,3-mixed glucans barley β-glucan and lichenan, the β1,4-glucan hydroxyethylcellulose (HEC), konjac glucomannan, oat spelt xylan, the β1,3-glucans laminarin and curdlan, carob galactomannan, potato galactan, pullulan or pustulan (data not shown). ITC confirmed that, outwith β1,3-glucans, CBM6-1 does not interact with β1,3-β1,4- mixed linked glucans or the other polysaccharides tested. In addition, CBM6-1 lacks the capacity to interact with insoluble cellulose preparations of Avicel and acid-swollen cellulose (data not shown).



C

Protein	Architecture	K_a (c^{-1})[a]
CBM6-2		21
CBM6-2/2		460
CBM6-1/2		7,7
B1 mutant	*	11,9
B12 mutant	* *	1,3
B2 mutant	*	0,4

[a] ***c*, % (w/v) ligand concentration**

Figure 4.3 Interaction of *Cm*Cel5B derivatives with lichenan analysed by affinity gel electrophoresis (AGE).

Recombinant proteins were purified through IMAC (Panel A) and subjected to AGE as described in Materials and Methods (Panel B). The proteins under analysis were CBM6-1/2 (lane 2), CBM6-1/2 with a mutation in cleft B of CBM6-1 (lane 3), CBM6-1/2 with a mutation in cleft B of CBM6-2 (lane 4), CBM6-1/2, with mutations in cleft Bs of both CBM6s (lane 5), CBM6-2 (lane 6) and CBM6-2/2 (lane 7). Lanes 1 and 8 contain protein molecular mass markers (Panel A) and BSA (Panel B).The affinity of *Cm*Cel5B truncated derivatives from Panels A and B for lichenan was determined through AGE (Panel C).

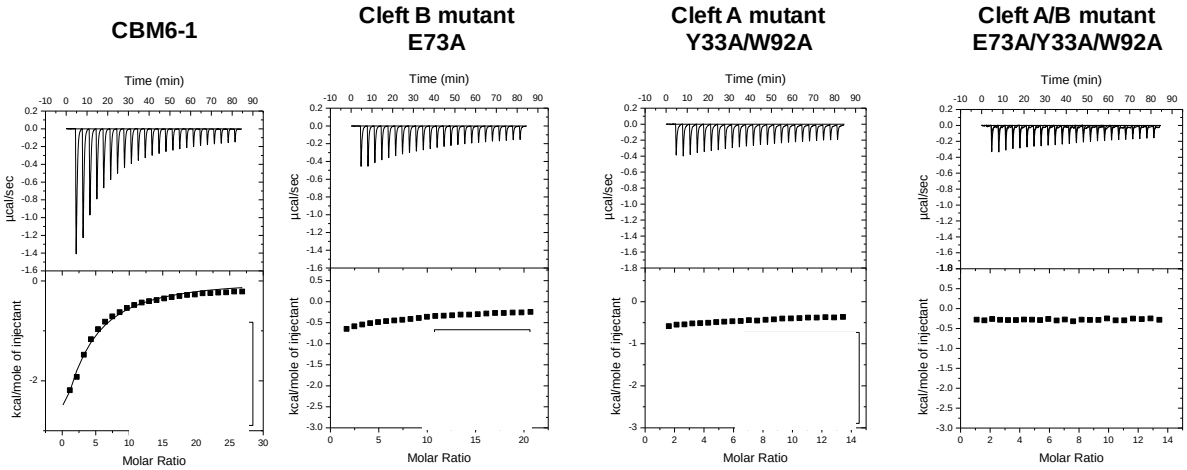


Figure 4.4 Isothermal titration calorimetry of wild type (CBM6-1) and mutant derivatives of CBM6-1 with laminarin.
Upper parts of each panel show the raw heats of binding and the lower parts are the integrated heats.

4.3.2. CBM6-1 and CBM6-2 do not act cooperatively to bind polysaccharides

Although CBM6-2 displays specificity for β1,4- or mixed linked β1,4-β1,3-glucans, reflecting the substrate specificity of the appended *Cm*Cel5B GH5 catalytic module (see below), while CBM6-1 recognizes β1,3-glucans, it is possible that the two modules display cooperativity particularly for polysaccharides that contain both β1,4 and β1,3-gluco linkages. To assess this possibility the capacity of CBM6-1 fused to CBM6-2 (CBM6-1/2) to bind to β-glucans was evaluated through AGE. The data, examples of which are presented in Figure 4.3B and the K_a values reported in Figure 4.3C, suggest that affinity of CBM6-1/2 for barley β-glucan is similar to CBM6-2, indicating that the two modules do not act co-operatively. To confirm that CBM6-2 can participate in avidity-mediated increases in affinity, two CBM6-2 modules were fused in tandem and the capacity of the bi-modular protein to interact with mixed linked β-glucan was evaluated. The data (Figure 4.3), confirms that a significant increase in affinity is observed when two copies of CBM6-2 are fused together. This phenomenon has been described for a variety of CBMs that display the same ligand specificities and are located in the same enzyme (Gilbert *et al.*, 2002; Bolam *et al.*, 2001; Boraston *et al.*, 2002). To further confirm the inability of CBM6-1 to affect the interaction of CBM6-2 with mixed linked β-glucan, the pivotal residues involved in ligand binding were mutated separately and together in the cleft B regions of the two modules within CBM6-1/2 and the affinity of the resulting proteins for mixed-linked β-glucan was

evaluated (Cleft B exclusively interacts with this polysaccharide in CBM6-2). The data (Figure 4.3) confirms that apparently only cleft B from CBM6-2 has the capacity to interact with mixed linked glucans. Therefore, taken together, the data presented here suggest that CBM6-1 does not act co-operatively with CBM6-2 to bind polysaccharides.

4.3.3. Mapping the ligand binding site by mutagenesis

To evaluate which CBM6-1 binding cleft interacts with laminarin, the affinity of various mutant derivatives was evaluated by ITC. By analogy with CBM6-2, Tyr33 and Trp92 are key residues involved in ligand recognition in cleft A, while Glu73 plays a major role in carbohydrate recognition in cleft B. Thus the following CBM6-1 mutants were constructed: Y33A (cleft A), W92A (cleft A), Y33A/W92A (cleft A), E73A (cleft B) and Y33A/E73A/W92A (clefts A and B) – CBM6-2 numbering used for simplification. The data, presented in Figure 4.4, revealed that all CBM6-1 mutant derivatives analysed display no significant binding affinity for laminarin, suggesting that both clefts contribute to laminarin recognition in CBM6-1. In addition, wild type CBM6-1 is unable to recognize both laminarihexose and cellohexaose (data not shown). Lack of affinity for laminarihexose suggests that CBM6-1 laminarin-binding cleft accommodates carbohydrates with a degree of polymerization higher than 6. Taken together these results indicate that CBM6-1 binds to laminarin through residues located both in clefts A and B. Therefore, family 6 CBMs have evolved the capacity to interact with laminarin in at least three different binding sites. Thus cleft A of *Cm*Cel5B CBM6-2 binds to the non-reducing end of all β-linked glucans or xylans assessed. In contrast, cleft A of *B. halodurans* CBM6 binds specifically to the non-reducing end of laminarin and the carbohydrate makes several productive interactions along its chain with the surface of the protein that is formed by the loops connecting the two β-sheets, which display a convex topology (Figure 4.5). The residues mediating ligand recognition in *Bh*CBM6, which are not located in cleft B, are not conserved in CBM6-1. In this work *Cm*Cel5B CBM6-1 is shown to accommodate laminarin through residues located in clefts A and B, revealing a third location for laminarin recognition within the CBM6 family. The discrete, and topologically distinct, laminarin binding sites may be intimately related with the capacity of CBM6 members to display an extensive range of ligand specificities and is in sharp contrast with other CBM families where variation in carbohydrate recognition reflects differences in the topology of a single binding site.

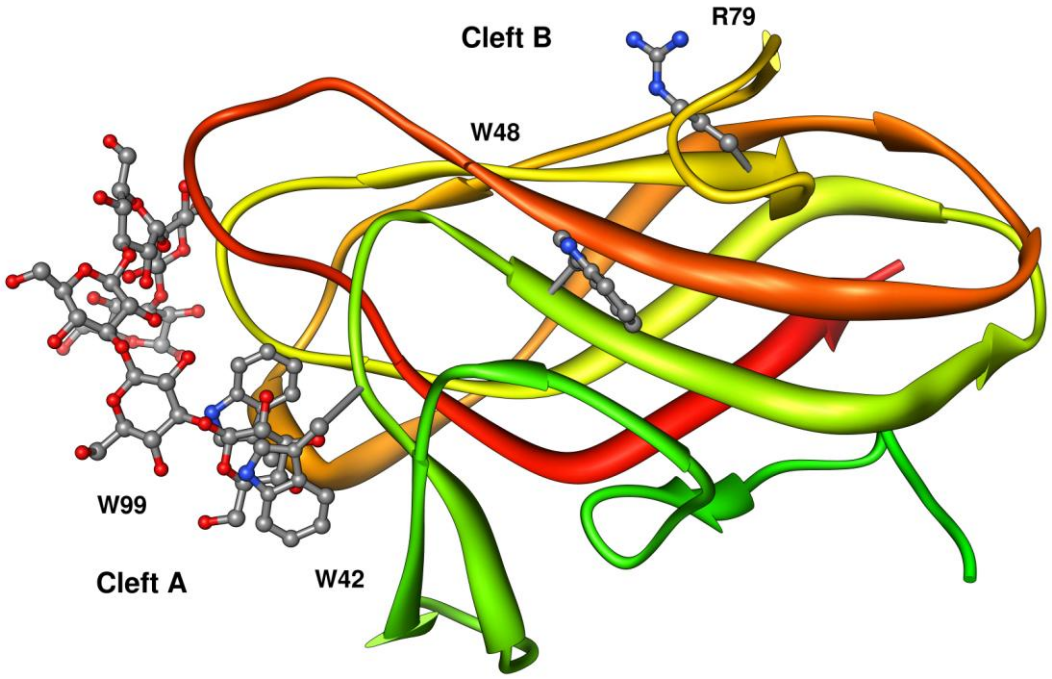


Figure 4.5 The three-dimensional structure of *Bh*CBM6 (PDB ID: 1W9W).
The three-dimensional structure of BhCBM6 (PDB ID: 1W9W). The structure of BhCBM6 is shown in color ramped format in complex with laminohexaose, which is depicted as grey carbons and red oxygens. Residues making contacts with the ligand at cleft A are highlighted. The figure demonstrates that laminarin does not interact with cleft B, which putative interacting residues are also depicted. The structure also shows that only a β1,3-glucan considerably longer than laminohexaose would be able to interact with both cleft A and B.

4.3.4. Role of CBM6 modules in the function of *Cm*Cel5B

To evaluate the role of the tandem CBM6s in the function of the glycoside hydrolase *Cm*Cel5B, derivatives of the enzyme, comprising the GH5 catalytic module alone (GH5) or fused to CBM6-1 and CBM6-2 (GH5-CBM6-1/2) were produced in the recombinant form. The temperature optimum of GH5 and GH5-CBM6-1/2 is ~37 °C with maximal activity at pH 7.5 (data not shown). The biochemical properties of both recombinant derivatives indicate that *Cm*Cel5B hydrolyses cellulosic substrates such as carboxymethyl cellulose, mixed linked β-1,4-β-1,3-glucans and glucomannans (Table 4.2). The enzyme did not hydrolyse laminarin, xyloglucan, mannan, galactomannan, carob-galactomannan, arabinogalactan, lupin galactan, arabinan, rhamnogalacturan, pectin galactan lupin, pectic galactan potato, polygalacturonic acid, rhamnogalacturan I, pustulan, pululan, pectin from apples and pectin from citrus. Comparison of the activities of GH5 and GH5-CBM6-1/2 revealed that the tandem CBM6s appear to

mediate a three-fold increase in the activity of *Cm*Cel5B against the insoluble substrate Avicel but not, significantly, against purified soluble polysaccharides (Table 4.2). This feature is common to many cellulose-binding CBMs (Fernandes *et al.*, 1999; Bolam *et al.*, 1998; Hall *et al.*, 1995) and is believed to result from the increased concentration of the enzyme on the surface of the recalcitrant substrate. As soluble purified polysaccharides are highly accessible to enzyme attack there is no requirement to sequester these biocatalysts on the surface of these isolated glycan chains.

Table 4.2 Specific activities of GH5 and GH5-CBM6-1/2 for a range of plant cell wall polysaccharides and targeting effect of CBM6-1/2

Substrate	Enzyme activity (kat[a])	
	GH5	GH5-CBM6-1/2
Specific activity[b]		
Avicel	10	30
β-glucan	19809	19694
Lichenan	18876	18575
Carboxymethylcellulose	14912	14779
Glucomannan	12267	12273
Xyloglucan (XG)	0	0

[a] mol of product formed per mol of enzyme min^{-1}.

[b] Assays performed at 37 ºC with 0.15% (w/v) of substrate except for Avicel where 2% (w/v) substrate was used.

[c] The enzymes do not present measurable reducing sugar activity against laminarin, xyloglucan, mannan, galactomannan, carob-galactomannan, arabinogalactan, lupin galactan, arabinan, rhamnogalacturan, pectin galactan lupin, pectic galactan potato, polygalactutronic acid, rhamnogalacturan I, pustulan, pululan, pectin from apples and pectin from citrus.

4.4. CONCLUSION: Biological rational for β1,3-glucan recognition by *Cm*Cel5B

In this report CBM6-1 from *C. mixtus* *Cm*Cel5B was shown to bind β1,3-glucans through residues located both in clefts A and B and the ligand binding interface is suggested to accommodate more than six β-1,3-linked glucose residues. The biological significance of this specificity is intriguing since β1,3-glucans are usually not present in the cell walls of terrestrial plants. However, it is well known that β1,3 polysaccharides are important components of fungal cell walls (Cid *et al.*, 1995). In addition, although β1,3-glucans are not known to be part of the cell walls of green, red or brown algae (Kloareg & Quatrano, 1988), laminarin that is localized in the vacuole is a storage polysaccharide in brown algae (Read *et al.*, 1996). *C. mixtus* is a saprophytic bacterium that colonizes the soil and, therefore, it is unlikely that algae polysaccharides are widely used as energy and carbon sources by this bacterium. However, it is possible

that *C. mixtus* hydrolyses fungal cell wall polysaccharides. As fungi are the initial colonizers of plant biomass, it is possible that saprophytic bacteria could have developed fungal lytic mechanisms allowing them to compete with fungal species that also use plant cell wall carbohydrates as an energy source. Intriguingly, the catalytic module appended to CBM6-1 consists of a typical cellulase arguing against this possibility. The fungal-targeting hypothesis may, however, be relevant to other *C. mixtus* glycoside hydrolases, such as *Cm*Cel16A, which is encoded by a gene that cluster with the *Cm*Cel5B locus (Centeno *et al.*, 2006). *Cm*Cel16A presents a GH16 catalytic domain, which typically hydrolyses β1,3-glucans, and a family 32 CBM that specifically recognizes β1,3-glucans (C. Fontes unpublished data). In contrast, CBM6-1 may target *Cm*Cel5B to plant cell wall locations that are being actively degraded by fungi where a variety of complex carbohydrates are accessible to enzyme attack. When the enzyme is located in the vicinity of the fungal cell wall, the action of the microbial eukaryote may expose plant cell wall β1,3-1,4-glucans, which are targeted by the C-terminal CBM6-2 of the glucanase, enabling the catalytic module to degrade its target substrate. Therefore, the cooperative action of CBM6-1 and CBM6-2 allows *Cm*Cel5B to be directed to the region of the cell wall that are actively being degraded by fungi, through CBM6-1, where the catalytic module, in harness with CBM6-2, will hydrolyse plant cell wall β1,3-1,4-glucans.

In addition to the above described hypothesis, CBM6-1 may also target the appended catalytic domain to the cell wall of oomycetes, a group of filamentous, unicellular heterokonts, physically resembling fungi, which may be plant pathogens (Bartnicki-Garcia, 1968). Nevertheless, this possibility seems unlikely since *C. mixtus* is not a plant pathogen itself. However, the CBM6-1 of *Cm*Cel5B may direct the enzyme to callose, a β-1,3-glucan molecule produce by plants in response to a fungal attack (Verma & Hong, 2001). This targeting mechanism would allow the enzyme to be in close proximity with the plant cell wall mixed linked glucans, which are the substrate for the enzyme catalytic domain. Finally, it is also possible that the primary substrates targeted by *Cm*Cel5B are the β1,3-1,4-glucans located in the fungal cell wall itself. It has been recently demonstrated that the cell wall of *Aspergillus fumigatus* contains linear β1,3-1,4-glucans, which corresponds to 10% of the total beta-glucan content (Fontaine *et al.,* 2002). Chitin, galactomannan and the linear β1,3-1,4-glucans were found to be covalently linked to the non-reducing end of β1,3-glucans, explaining the

specificity of CBM6-1 (Fontaine *et al.*, 2002). The prevalence of β1,3-1,4-glucans in the cell wall of cellulolytic and hemicellulolytic fungi, however, is currently unknown.

CHAPTER 5 A novel non-catalytic Carbohydrate-Binding Module (CBM) displays specificity for galactose-containing polysaccharides.

Márcia A.S. Correia[1], Cedric Y. Montanier[2], James E. Flint[2], José A.M. Prates[1], Susan Faribanks[2], Richard J. Lewis[2], Pedro M. Coutinho[3], Carlos M.G.A. Fontes[1] and Harry J. Gilbert[2]

[1]CIISA - Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [2]Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom; [3]Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités Aix-Marseille I & II, 163 Avenue de Luminy, 13288 Marseille, France;

submitted for publication

Abstract

The enzymic deconstruction of plant cell walls plays a central role in the carbon cycle and is of increasing environmental and industrial significance, as this process is an important component of lignocellulosic-based biofuel production. Enzymes that catalyse this process generally display a complex molecular architecture in which the catalytic modules are appended to non-catalytic carbohydrate-binding modules (CBMs). CBMs potentiate the rate of catalysis by bring their cognate enzymes into intimate contact with the target substrate. One of the most powerful plant cell wall degrading systems is the multienzyme complex, termed the "cellulosome", synthesised by the bacterium *Clostridium thermocellum.* The cellulosome scaffolding protein contains a CBM that binds to crystalline cellulose, which anchors the complex onto the plant cell wall, while the individual catalytic subunits direct these macromolecular structures to regions of the wall that contain the target substrates for these enzymes. Here we identify a novel CBM (*CtCBM55*) within the large *C. thermocellum* cellulosomal protein Cthe_2193, which establishes a new CBM family. *CtCBM55*, in contrast to previously characterized cellulosomal CBMs, which recognize either xylans or β-glucans, binds to D-galactose and L-arabinose in either anomeric configuration. The ligand binding site in the β-sandwich protein is located in the loops that connect the two β-sheets. Specificity is conferred through numerous interactions with the axial *O4* of the target sugars, a feature that distinguishes galactose and arabinose from the other major sugars located in plant cell walls. *CtCBM55* displays tighter affinity for multivalent ligands compared to molecules containing single galactose residues, which is associated with precipitation of these complex carbohydrates. These avidity effects, which confer the targeting of polysaccharides, are mediated by calcium-dependent oligomerization of the CBM. The significance of the novel specificity and the oligomeric structure displayed by *CtCBM55* within the context of a multienzyme complex is discussed.

5.1. INTRODUCTION

Carbohydrate protein recognition plays a central role in biology exemplified by microbial-mediated plant cell wall degradation. The release of sugars from plant cell walls is not only critical to the maintenance of the carbon cycle, but is of increasing industrial and environmental significance through the generation of second generation lignoceullulosic-based biofuels (Ragauskas *et al.*, 2006). The chemical and physical complexity of plant cell walls restricts their accessibility to enzyme attack and thus the recycling of photosynthetically fixed carbon is a relatively slow biological process.

Enzymes that catalyse plant cell wall degradation display complex molecular architectures. In addition to the catalytic module(s), these enzymes often contain one or more non-catalytic carbohydrate-binding modules (CBMs) (Boraston, *et al.*, 2004). CBMs, by binding to their target ligand, bring the appended enzymes into intimate contact with its substrate thereby potentiating catalysis by reducing the accessibility problem (Black *et al.*, 1997; Bolam *et al.*, 1998; Ferreira *et al.*, 1990). Carbohydrate modifying enzymes and their component modules, which include CBMs, have been classified into sequence-based families in the CAZy database (Cantarel *et al.*, 2009). Currently there are 54 families of CBMs (June 2009), which recognize a variety of plant and mammalian glycans. In some families, exemplified by CBM1 and CBM3, ligand specificity is invariant; however, it is becoming increasingly apparent that carbohydrate recognition can be highly variable in different members of the same family. For example, CBM6 contains members that recognise xylan (β-1,4-xylose polymer), single cellulose β-1,4-glucose polymer) chains, mixed linked β-1,4-β-1,3-glucans, laminarin β-1,3-glucose polymer) and agarobiose (3,6-anhydro-L-galactose-D-galactose) (Czjzek *et al.*, 2001; Henshaw *et al.*, 2006; Henshaw *et al.*, 2004; van Bueren *et al.*, 2005).

Representative three dimensional structures of around half the CBM families are available. The majority of these proteins display a β-sandwich fold in which the ligand binding site for extended glycan chans is generally located on the β-sheet that comprises the concave surface of the protein (see Boraston *et al.*, 2004 for review). For CBMs that recognize terminal mono- or disaccharides, the ligand binding site is often located in the loops that connect the two β-sheets. While there are examples of CBMs where the orientation of aromatic residues confers ligand specificity (Raghothama *et al.*, 2000; Simpson *et al.*, 2000), it is increasingly evident that polar interactions play a more important role in carbohydrate recognition (Henshaw *et al.*, 2006; Henshaw *et al.*, 2004; Pell *et al.*, 2003). Recent reports have also revealed examples of CBMs that

harness calcium in ligand recognition (Henshaw *et al.*, 2006; Jamal-Talabani *et al.*, 2004; Montanier *et al.*, 2009). CBMs are located exclusively in monomeric enzymes, thus co-operative (or avidity) effects occur rarely. Exceptions to this general rule are, however, evident in enzymes that contain multiple copies of CBMs that display the same specificity, where avidity effects between these modules has led to increased affinity for polysaccharides such as xylan (Bolam *et al.*, 2001; Boraston *et al.*, 2002a; Charnock *et al.*, 2002). Given that oligomerization is a general feature of lectins (carbohydrate-binding modules that are not components of enzymes that modify carbohydrates), where avidity effects are common (Vijayan *et al.*, 1999), it is surprising that similar macromolecular associations have not been observed in the CBM literature.

The multienzyme complex expressed by *Clostridium thermocellum*, referred to as the "cellulosome", is one of the most powerful and intricate plant cell wall degrading system described (Bayer *et al.*, 2004). The enzyme complex is anchored onto the plant cell wall by the non-catalytic scaffoldin protein, which contains a family 3 CBM that binds to crystalline cellulose (Tormo *et al.*, 1996). The catalytic subunits of the cellulosome also contain CBMs, which target β-glucans, xylans, uronic acids, chitin and pustulan (Carvalho *et al.*, 2004 ; Charnock *et al.*, 2000; Czjzek *et al.*, 2001; Dvortsov *et al.*, 2009; Montanier *et al.*, 2009; Najmudin *et al.*, 2006) and thus direct the complex towards the target substrates for these enzymes. The extent to which the cellulosome can be directed towards the myriad of carbohydrate structures in plant cell wall, however, remains relatively unexplored and, indeed, numerous cellulosomal proteins contain modules of unknown function that could, potentially, comprise CBMs that display novel specificities. Furthermore the quaternary structure of the cellulosome may provide a background for cooperativity between the CBMs within this protein complex, although such interactions have not been observed experimentally.

Here we report the structure and biochemistry of a module, defined as *Ct*CBM55, of the *C. thermocellum* cellulosomal protein Cthe_2193 (GenBank protein accession ABN53395.1) that exhibits no significant sequence similarity to CBMs in the CAZy database. *Ct*CBM55 displays a β-sandwich fold that binds to terminal α- and β-D-galactopyranose or L-arabinopyranose residues of complex polysaccharides. Ligand specificity is conferred primarily through extensive interactions with the axial *O4* of galactose and arabinose, a distinctive feature of these two sugars. *Ct*CBM55 also displays calcium-mediated oligomerization resulting in avidity effects that confer selectivity for polysaccharides rather than monovalent oligosaccharides.

5.2. EXPERIMENTAL PROCEDURES

5.2.1. Cloning, expression and purification of components of Cthe_2193

An initial analysis of the *Clostridium thermocellum* genome was performed as described in Cantarel *et al.* (2009), revealing the modularity of proteins corresponding to all identifiable elements corresponding to CAZy families of catalytic and carbohydrates-binding modules (see http://www.cazy.org). This analysis was complemented with the identification of cellulosomal modules corresponding to dockerins and cohesins (data not shown). DNA encoding the following regions of Cthe_2193, GH5, GH5-CBM6, CBM6-CBM13-Fn3-*Ct*CBM55 and *Ct*CBM55 were amplified using primers from Table 5.1, containing NheI and XhoI restriction sites, listed in Supplementary Information. The amplified DNAs were cloned into NheI/XhoI restricted pET21a, such that the encoded recombinant proteins contain a C-terminal His_6 tag. To express the four *C. thermocellum* proteins *Escherichia coli* strain BL21(DE3), harbouring appropriate recombinant plasmids, was cultured to mid-exponential phase in Luria broth at 37 ^{o}C followed by the addition of isopropyl β-D-galactopyranoside at 1 mM, to induce recombinant gene expression, and incubation for a further five hours at 37 ^{o}C. The recombinant proteins were purified to >90 % electrophoretic purity by immobilized metal ion affinity chromatography using Talon^{TM}, a cobalt-based matrix, and elution with 100 mM imidazole, as described previously (Charnock *et al.*, 2002). When preparing the seleno-methionine derivatives of *Ct*CBM55 and GH5-CBM6 for crystallography, the proteins were expressed in *E. coli* B834 (DE3), a methionine auxotroph, cultured in media comprising 1 litre SelenoMet Medium Base^{TM}, 50 ml SelenoMet Nutrient Mix^{TM} and 4 ml Selenomethionine solution (10 mg/ml). Recombinant gene expression was as described above as was protein purification except that all buffers were supplemented with 10 mM β-mercaptoethanol.

5.2.2. Mutagenesis

Site-directed mutagenesis was carried out using PCR-based QuikChange method (Stratagene) deploying the primers listed in Supplementary Information.

Table 5.1 Primers used for cloning components of Cthe_2193 and for the mutagenesis of CtCBM55

Protein	Primers	Sequence (5'-3')
*Ct*CBM55	Forward	CTCGCTAGCTATCCTAAACTTACGG
	Reverse	CACCTCGAGATGCACATCATCATTC
GH5	Forward	CTCGCTAGCAGCCCGCAACGTGGCCGG
	Reverse	CACCTCGAGGTCATAATACGAAACCCC
GH5-CBM6	Forward	CTCGCTAGCCCGCAACGTGGCCGG
	Reverse	CACCTCGAGTATCGGAGAAAGTTC
CBM6-CBM13-Fn3-*Ct*CBM55	Forward	CTCGCTAGCACGGATTCGGTGAATG
	Reverse	CACCTCGAGATGCACATCATCATTC
W18A	Forward	GGAACCCAAGGTTCGGCGAATAACATTGGG
	Reverse	CCCAATGTTATTCGCCGAACCTTGGGTTCC
F36A	Forward	GGTGACCTGAACACGGCTTTTGACGGTCCT
	Reverse	AGGACCGTCAAAAGCCGTGTTCAGGTCACC
D38A	Forward	GAACACGTTTTTTGCCGGGCCCACAGCAAACGG
	Reverse	CCGTTTGCTGTGGGCCCGGCAAAAAATGTGTTC
W46A	Forward	CAGCAAACGGCTGCGCGCTGGGACTGGATTTTG
	Reverse	CAAAATCCAGTCCCAGCGCGCAGCCGTTTGCTG
D50A	Forward	TGGCTGGGACTGGCTTTTGGGGAAGGTGTG
	Reverse	CACACCTTCCCCAAAAGCCAGTCCCAGCCA
E53A	Forward	GGACTGGATTTTGGGGCAGGTGTGAGGAATGTC
	Reverse	GACATTCCTCACACCTGCCCCAAAATCCAGTCC
R67A	Forward	ATTAAATTCTGCCCGGCGTCCGGCTATGAACAG
	Reverse	CTGTTCATAGCCGGACGCCGGGCAGAATTTAAT
Y70A	Forward	GCCCGCGTTCCGGCTATGAACAGCGCATGATAG
	Reverse	CTATCATGCGCTGTTCATAGCCGGAACGCGGGC
R73A	Forward	CCGGCTATGAACAGCGCATGATAGGGGGAATT
	Reverse	AATTCCCCCTGTCATGCGCTGTTCATAGCCGG
Q80A	Forward	GGGGGAATTTTTGCGGGGGCAAATAAAG
	Reverse	CTTTATTTGCCCCCGCAAAAATTCCCCC
E85A	Forward	GGGGCAAATAAAGCAGAATTCAGCGATGC
	Reverse	GCATCGCTGAATTCTGCTTTATTTGCCCC
F87A	Forward	GCAAATAAAGAAGATGCTAGCGATGCAGTG
	Reverse	CACTGCATCGCTGGCATCTTCTTTATTTGC
R120A	Forward	TTCCGCTATGTCGCGTATTTGTCCCCGGAC
	Reverse	GTCCGGGGACAAATACGCGACATAGCGGAA
N128A	Forward	CCGGACGGCAGTGCTGGAAATATTGCA
	Reverse	TGCAATATTTCCAGCACTGCCGTCCGG

5.2.3. Assays

GH5 and GH5-CBM6 were assayed for enzyme activity using the method of Miller to detect the presence of reducing sugar. The standard assay was carried out in 50 mM sodium phosphate buffer, pH 7.0, and the potential polysaccharide substrate at 10 mg/ml. The reactions were initiated by the addition of potential enzyme up to 50 µM and incubation at 55 ^{o}C for up to 16 h. The identification of potential reaction products were also assessed by HPAEC using methodology described previously (Charnock *et al.*, 1997). The capacity of GH5 and GH5-CBM6 to hydrolyse xylooligosaccharides was assessed by HPAEC using 100 µM of oligosaccharide and 50 µM of protein. Reactions were incubated at 55 ^{o}C for up to 24 h.

Affinity gel electrophoresis was used to screen for the binding of *Ct*CBM55 to polysaccharides, following the method of (Freelove *et al.*, 2001). The proteins were subjected to non-denature PAGE, in the presence of 5 mM $CaCl_2$, deploying parallel gels containing no ligand and the target polysaccharide at 100 µg/ml, respectively. The gels were also loaded with BSA which acts as a non-binding negative control. After electrophoresis the gels were stained with Coomassie Blue and proteins that bound to the polysaccharide displayed reduced electrophorestic migration in the presence of the complex carbohydrate. The binding of *CtCBM55* was quantified by isothermal titration calorimetry, as described previously (Charnock *et al.*, 2000). Titrations were carried out in 50 mM Na/Hepes buffer, pH 7.5, containing 5 mM $CaCl_2$ (or 5 mM EDTA) at 25 ^{o}C. The reaction cell contained protein at 145 µM, while the syringe contained either the oligosaccharide at 10 mM or polysaccharide at 3-5 mg/ml. The titrations were analyzed using Microcal Origin version 7.0 software to derive, n, K_a and ΔH values, while ΔS was calaculated using the standard thermodynamic equation:

$$-RT\ln K_a = \Delta G = \Delta H - T\Delta S.$$

5.2.4. Crystallography

Proteins were purified by immobilized metal ion affinity chromatography (IMAC) using Talon™ resin and elution in 20 mM Tris–HCl buffer (pH 8.0) containing 300 mM NaCl and 100 mM imidazole. The eluted proteins were then dialyzed against buffer A (20 mM Tris–HCl, pH 8.0) and applied to a Bio-Rad Q12 column. Proteins were crystallized using the hanging drop vapour technique at 20 ^{o}C with an equal volume (1 µl) of protein and reservoir solution using 1 µl of protein. The seleno-methionine derivative apo form of *Ct*CBM55 (in the presence of 5 mM DTT) and the native form of the protein (161 mg/ml) in complex with either 10 mM xyloglucan oligosaccharides or 33 mM L-arabinose were crystallized in 20 % (w/v) PEG 3350, 0.2 M tri-sodium citrate (PEG/Ion^{TM} 1 screen condition 46; Hampton Research). Native *Ct*CBM55 (191 mg/ml) in complex with ^{1}6-α-GalMan$_3$ (10 mM) was crystallized in 0.5 M ammonium sulphate in 0.1 M NaHepes buffer, pH 7.4., containing 30% 2-methyl-2,4-pentanediol (condition 59 of the PACT screen). Crystals, which grew in 4-5 days, were stabilized by adding 1 µl cryoprotectant solutions (30% (v/v) glycerol in the crystallization buffer, stepwise over a few days for equilibration, and flash-frozen in

liquid nitrogen. Data were collected using a Rigaku 007 MicroMax copper rotating anode using an RAXIS IV++ imaging plate detector to 1.7 Å resolution (except *CtCBM55*-arabinose, which was collected to 2.0 Å). Crystals of all three proteins were in the space group $F2_3$. The cell dimensions of the crystals were $a = b = c = 191.3$ Å, $\alpha = \beta = \gamma = 90^{o}$, with two molecules of *CtCBM55* in the asymmetric unit. The structure of the SelMet apo form of *CtCBM55* was solved by SAD (Evans *et al.*, 2002). The structures of the other forms of the CBM were determined by molecular replacement with MOLREP (Vagin *et al.*, 2000) using apo-CtCBM55 as the search model. The structure was refined using REFMAC (Lamzin *et al.*, 1997) with final manual corrections using COOT (Emsley *et al.*, 2004). The data collection, phasing and refinement statistics are displayed in Table 5.2.

Table 5.2 Crystal parameters and refinement statistics of native *Ct*CBM55 and in complexes with ligands

	Native	**Xyloglucan oligosaccharide**	**GM3**	**A2**
Data collection				
Space group	$F2_3$	$F2_3$	$F2_3$	$F2_3$
Cell dimensions				
$a = b = c$ (Å)	191,39	191.69	191.35	191.73
$\alpha = \beta = \gamma$ (°)	90	90	90	90
Resolution (Å)	22.22-1.70 (1.79-1.70)	47.25-1.75 (1.80-1.75)	110.4-1.70 (1.74-1.70)	110.4-2.00 (2.05-2.00)
R_{merge}	0.182 (0.528)	0,174 (0.965)	0.146 (0.895)	0.121 (0.736)
$I / \sigma I$	10.7 (2.1)	17.0 (3.1)	14.0 (1.8)	16.9 (1.8)
Completeness (%)	100.0 (100.0)	100.0 (100.0)	99,96 (99,72)	99.72 (98.83)
Redundancy	8.3 (8.2)	15.4 (15.1)	7.4 (7.2)	5.9 (5.0)
Refinement				
No. reflections	64559	58558	60268	37425
R_{work} / R_{free}	0.189/0.204	0.172/0.193	0.175/0.194	0.197/0.216
No. atoms				
Protein	2254	2199	2165	2118
Ligand / Ion	0/2	52/2	37/2	20/2
Water	480	416	503	135
R.m.s.[a] deviations				
Bond lengths (Å)	0.012	0.024	0.031	0.016
Bond angles (°)	1.49	1.13	2.4	1.38

[a]R.m.s .: Root mean square.

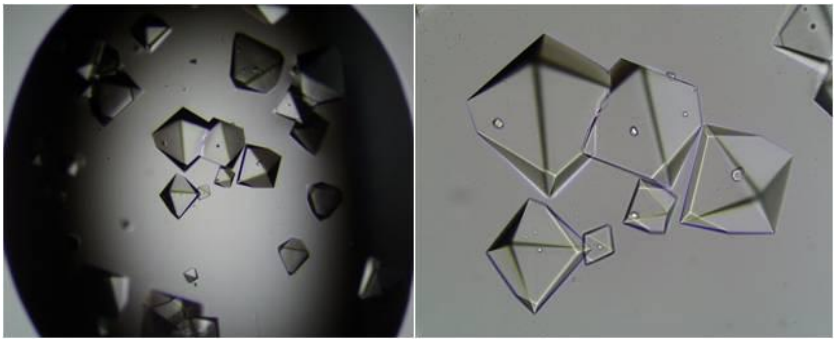


Figure 5.1 Crystals of *Ct*CBM55.
The Crystals were obtained by hanging-drop vapor diffusion in the presence of 0.5 M ammonium sulphate in 0.1 M NaHepes buffer, pH 7.4., containing 30% 2-methyl-2,4-pentanediol.

5.3. RESULTS

5.3.1. *Ct*CBM55 is a component of the *C. thermocellum* cellulosomal protein Cthe_2193.

Inspection of the genome sequence of *C. thermocellum* reveals 72 genes encoding putative proteins containing type I dockerins, revealing the repertoire of components of the bacterium cellulosome, a multienzyme plant cell wall degrading complex. The function of many of these proteins have been demonstrated biochemically, or inferred from sequence similarity to proteins with known activities. There remain, however, several cellulosomal proteins that contain modules with no sequence similarity with proteins of known function. The open reading frame identified in locus Cthe_2193 is one such protein. This potentially secreted 948 residue polypeptide (contains an N-terminal signal peptide) consists of six modules, which are displayed in Figure 5.2. A Particularly noteworthy feature of the open reading frame is the sequence extending from 739 to 885, defined hereafter as *Ct*CBM55, which displays no significant sequence similarity to proteins in the CAZy database. To investigate the function of *Ct*CBM55, a recombinant form of this protein module was expressed in soluble form in *E. coli* and purified to electrophoretic homogeneity.

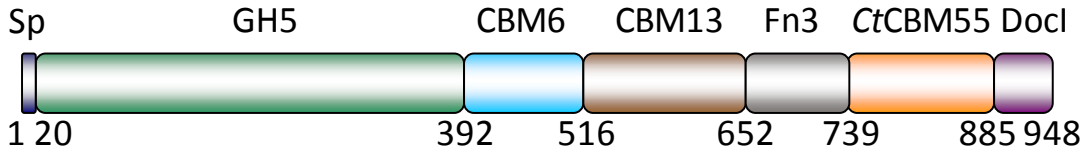


Figure 5.2 Schematic of Cthe_2193
The abbreviated modules are as follows: Sp, signal peptide; DocI, type I dockerin modules.

5.3.2. Biochemical properties of *Ct*CBM55

Biochemical analysis of *Ct*CBM55 failed to reveal catalytic activity against an extensive range of plant structural polysaccharides including cellulose and various hemicellulose and pectin polymers. To explore whether *Ct*CBM55 fulfils a non-catalytic carbohydrate binding function the capacity of the protein to bind to a range of polysaccharides was assessed by affinity gel electrophoresis. The data, reported in Table 5.3., show that the protein module binds to xyloglucan, arabinogalactan and galactomannan but does not recognize the other polymers evaluated. *Ct*CBM55 therefore comprises a non-catalytic carbohydrate-binding module and represents the founding member of a new CBM family designated CBM55. To investigate, in more detail, the specificity of *Ct*CBM55, the capacity of the protein to bind to galactomannan and xyloglucan constituents was assessed by isothermal titration calorimetry. The data, reported in Table 5.4 with example titrations displayed in Figure 5.3, show that *Ct*CBM55 does not bind to mannooligosaccharides or cellulooligosaccharides, the backbone structures of galactomannan and xyloglucan, respectively, or the repeating XXXG (X is a backbone glucose decorated α1,6 with Xyl and G represents unsubstituted glucose) motif of xyloglucan (Gloster *et al.*, 2007). The protein, however, does bind to 6^1-α-D-$GalMan_3$, Gal_2Man_5 and a mixture of XLXG, XLLG and XXLG, where L is X in which the xylose is decorated with an β1,2-linked D-galactose residue. The stoichiometry of binding was unity for galactose residues (in the oligosaccharides) indicating that *Ct*CBM55 has a single ligand binding site. The thermodynamic data show that the binding of *Ct*CBM55 to all its ligands is enthalpically driven, while the change in entropy makes a negative contribution to overall affinity, as observed for the majority of CBMs studied to date (Boraston *et al*, 2004; Boraston *et al*, 2002a; Carvalho *et al.*, 2004; Charnock *et al.*, 2000; Czjzek *et al.*, 2001; Henshaw *et al*, 2006; Henshaw *et al*, 2004; Montanier *et al.*, 2009; Najmudin *et al.*, 2006; Pell *et al.*, 2003). These data indicate that the capacity of *Ct*CBM55 to bind to galactomannan, arabinogalactan and xyloglucan is through its recognition of side chains containing terminal galactose residues, which are known to be present in these polysaccharides. This view was confirmed by the observation that the protein binds to galactose but not to sugars, such as xylose, mannose and glucose, where *O4* is equatorial. Indeed, the similar affinity of *Ct*CBM55 for 6^1-α-D-$GalMan_3$ and galactose, and the lack of significant binding to XXXG, suggests that the backbone component of mannan and xyloglucan do not make significant interactions with the protein. To summarize, the capacity of *Ct*CBM55 to

bind to xyloglucan and galactomannan is conferred by its recognition of either α- or β-D-galactose.

Table 5.3 Polysaccharide specificity of *Ct*CBM55 determined by affinity polyacrylamide gel electrophoresis

Ligand[a]	Binding[b]
Arabinogalactan	++
Oat-spelt Xylan	-
Wheat arabinoxylan	-
Laminarin	-
Lichenan (Icelandic moss)	-
Xyloglucan (Tamarind)	+++
Glucomannan (Konjac)	-
Galactomannan (Gal:Man, 21:79) (Carob)	++
Galactomannan (Gal:Man, 38:40) (Guar)	+++
Pullulan	-
Galactan (Potato)	-
Pustulan	-
Rhamnogalacturonan (Soya bean pectic fibre)	-
Hydroxyethylcellulose (HEC)	-
Glucuronoxylan	-
Arabinan (debranched)	-
Pectic galactan (Potato)	-
Polygalacturonic acid	-
Rhamnogalacturonan I (Potato)	-

[a]Ligands were screened at a concentration of 0.1 mg/ml.
[b] No detectable binding, -; significant binding, ++; strong binding +++.

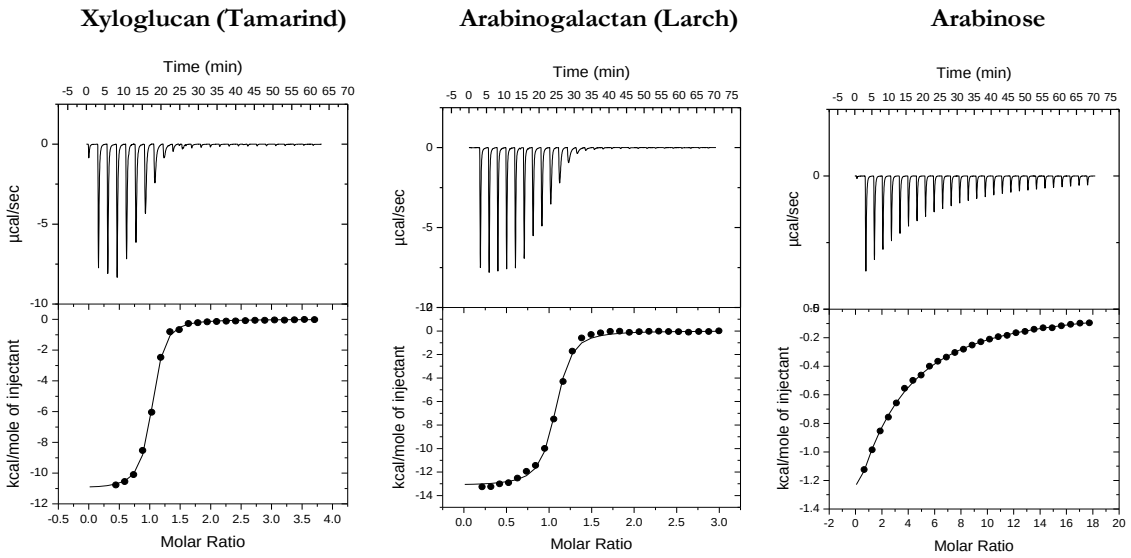


Figure 5.3 Representative ITC data of *Ct*CBM55 binding to soluble ligands.
The ligand (3 mg/ml xyloglucan, 5 mg/ml arabinogalactan, 10 mM arabinose) in the syringe was titrated into *CtCBM55* (145 µM) in the cell. The top half of each panel shows the raw ITC heats; the bottom half, the integrated peak areas fitted using a one single binding model by MicroCal Origin software. ITC was carried out in 50 mM Na/Hepes, pH 7.5, containing 5 mM $CaCl_2$, at 25 °C.

Table 5.4 Affinity and thermodynamic parameters of *Ct*CBM55 binding to polysaccharides, oligosaccharides and monosaccharides.

Data collected for titrations with 5 mg/ml of polysaccharide were fitted for xyloglucan, galactomannan, arabinogalactan and arabinan with a molar concentration of binding ligand of 3.1 mM, 4.1 mM, 2.1 mM and 3.2 mM respectively. Oat-spelt xylan was at 10 mg/ml. Xylooligosaccharide, xyloheptasaccharide, G_2M_5, GM_3, mannopentaose, cellohexaose and cellotetraose were at 3.9 mM, 3 mM, 3 mM, 2.4 mM, 3.1 mM, 3 mM and 3.1 mM respectively. Galactose and arabinose were at 10 mM and xylose at 20 mM. *Ct*CBM55 was at 145 µM in the cell.

Ligand	$K_a \times 10^4$ (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	n[a]
Galactomannan (Carob)	56.4 ± 3.1	-7.85 ± 0.03	-11.23 ± 0.06	-3.38 ± 0.03	1.00 ± 0.003
6^3, 6^4-di-α-D-galactosyl mannopentaose (G_2M_5)	2.21 ± 0.72	-4.56 ± 1.14	-8.69 ± 1.6	-4.13 ± 0.46	0.992 ± 0.15
6^1-α-D-galactosyl mannotriose (GM_3)	0.29 ± 0.02	-4.33 ± 0.03	-7.04 ± 0.75	-2.71 ± 0.72	1.07 ± 0.89
Galactose	0.32 ± 0.003	-4.79 ± 0.01	-8.59 ± 0.17	-3.8 ± 0.16	1.07 ± 0.02
Mannopentaose	n.b.[b]	-	-	-	-
Xyloglucan (Tamarind)	177 ± 17.7	-8.53 ± 0.06	-11.63 ± 0.07	-3.1 ± 0.01	1.01 ± 0.003
Xyloglucan oligosaccharide	2.01 ± 0.20	-5.87 ± 0.06	-10.74 ± 0.46	-4.87 ± 0.40	1.08 ± 0.036
Xyloglucan heptasaccharide	< 0.01	-	-	-	-
Arabinogalactan (Larch)	84.1 ± 8.4	-8.09 ± 0.06	-13.16 ± 0.10	-5.07 ± 0.04	1.04 ± 0.005
Arabinan (debranched)	9.42 ± 0.49	-6.79 ± 0.03	-10.84 ± 0.10	-4.05 ± 0.07	1.02 ± 0.007
Arabinose	0.11 ± 0.005	-4.17 ± 0.02	-9.10 ± 0.20	-4.93 ± 0.18	0.99 ± 0.22
Xylan (Oat-spelt)	< 0.01	-	-	-	-
Cellohexaose	n.b.	-	-	-	-
Cellotetraose	n.b.	-	-	-	-
xylose	n.b.	-	-	-	-

[a]The ITC data were fitted to a single site binding model for all ligands. For polysaccharide ligands in which the molar concentration of binding sites is unknown, the n-value was iteratively fitted to as close as possible to one, by adjusting the molar concentration of the ligand

[b]n.b.: no binding.

5.3.3. The crystal structure of *CtCBM55*

To explore the mechanism by which *Ct*CBM55 binds to both α- or β-D-galactose, the crystal structure of the protein was determined. The final structure encompasses residues 3-142 of the recombinant from of *Ct*CBM55, corresponding to residues 739-878 of full length Cthe_2193. The structure contains 525 water molecules. In common with the vast majority of CBM familes, *Ct*CBM55 displays a canonical β-sandwich fold (see Boraston *et al.*, 2004 for review) comprising two β-sheets containing five antiparallel β-strands on the concave face (β-1, 2, 4, 5 and 7) and three antiparallel β-strands on the convex face (β-3, 6 and 8). The two small α-helices and five loops on the top of the β-sandwich complete the structure, Figure 5.4. The protein contains a bound metal located at the beginning of the loop connecting the α-1 helix to β-2 which, based on its B-factor, has been modelled as calcium. Indeed this structural calcium is conserved in many CBM families, such as CBM4, CBM6, CBM29 and CBM32 (Boraston *et al.*, 2002b; Charnock *et al.*, 2002; Czjzek *et al.*, 2001; Ficko-Blean *et al.*,

2005). A structural homology search using the SSM server identified a large number of structurally similar proteins displaying a β-sandwich fold, including several CBM families. Significantly, the structure of the loop region comprising the ligand binding site, described in detail below, was not highly conserved in these proteins. Indeed, structural similarity to the ligand binding site of *Ct*CBM55 is evident in only the four proteins that display greatest structural homology with *C. thermocellum* protein; *Mv*CBM32 (PDB code 1EUT), *Cp*CBM32 (PDB code 2J1A), *Sp*CBM47 (PDB code 2J1R) and *Cd*CBM32 (PDB code 1GOF). These proteins display < 14 % sequence identity with *Ct*CBM55 and root mean square deviation values of 2.7 Å over 131 Cα atoms (*Mv*CBM32), 2.4 Å over 126 Cα atoms (*Cp*CBM32), 2.3 Å over 125 Cα atoms (*Sp*CBM47) and 2.7 Å over 131 Cα atoms (*Cd*CBM32).

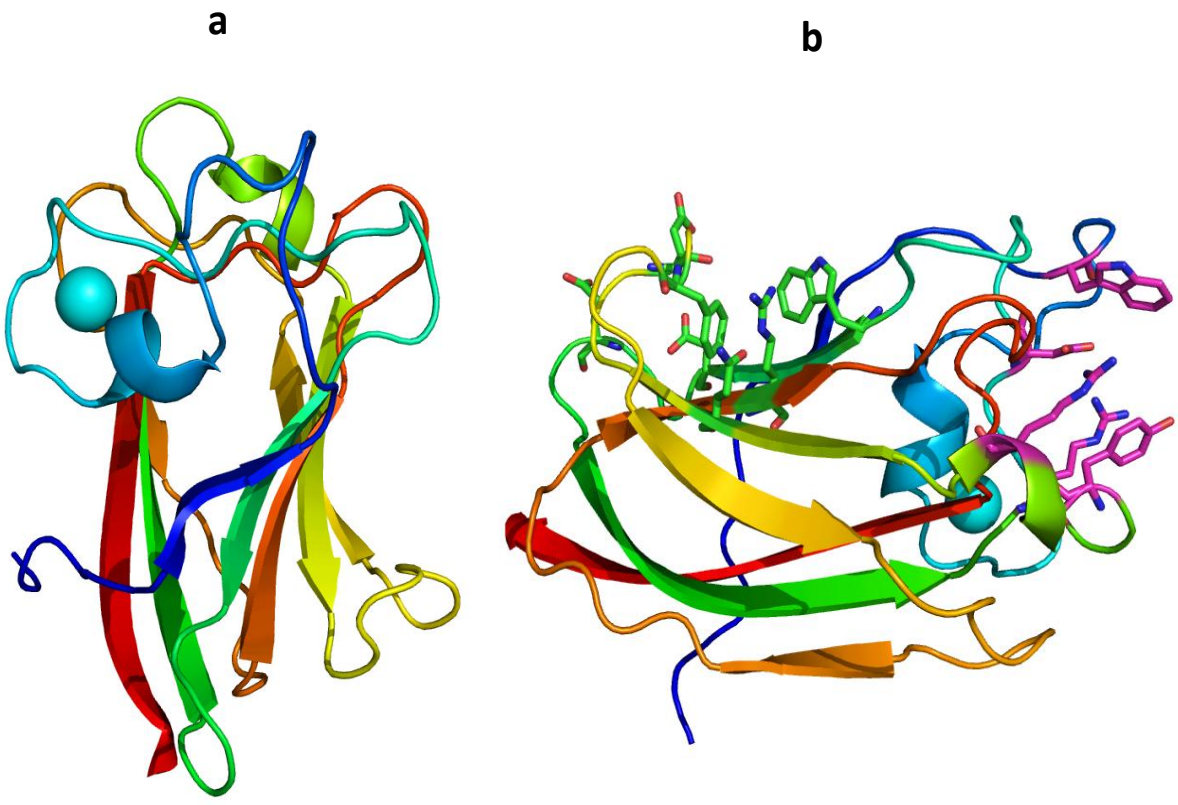


Figure 5.4 The crystal structure of *Ct*CBM55 and the location of residues that interact with galactose

The structure of *CtCBM55* (*Panel a* and *b*) is displayed in ribbon representation color-ramped from the **N**-terminus (blue) to the **C**-terminus (red). The residues on the concave surface of the protein and in the loops connecting the two β-sheets are in green and magenta, respectively. The residues highlighted on the concave surface are Trp-46, Asp-50, Glu-53, Gln-80, Glu-85, Phe-87 and Arg-120, while those in the loops are Trp-18, Asp38, Arg-67, Tyr-70 and Arg-73. The bound calcium is shown as a cyan sphere.

Inspection of the surface of the protein reveals a shallow pocket formed by the loops connecting β-1 and α-1, α-1 and β-2, β-2 and α-2, while α-2 also contributes to the binding pocket, Figure 5.5 and Figure 5.6, a topology consistent with the specificity of the protein for the terminal sugars of complex polysaccharides. The crystal structure of *Ct*CBM55 in complex with ^{1}6-α-D-GalMan$_3$ and XLXG confirms that the pocket does indeed represent the ligand binding site and provides insight into how the protein recognizes the galactosyl residue, which makes identical interactions with the protein in both ligand complexes. Thus, the α-face of the pyranose ring of the bound galactose makes extensive hydrophobic contacts with Trp-18 which is aligned parallel to the sugar ring. Such hydrophobic interactions are a generic feature of CBM ligand recognition (Charnock *et al.*, 2002; Nagy *et al.*, 1998; Pell *et al.*, 2003; Raghothama *et al.*, 2000; Simpson *et al.*, 2000). The bound galactose also makes several direct hydrogen bonds with the protein; *O1* and *O2* make polar contacts with the OH of Tyr-70, while *O3* forms hydrogen bonds with Nη1 and Nη2 of Arg-67 and Oδ2 of Asp-38. The *O4* of galactose makes numerous potential polar interactions with *Ct*CBM55; the hydroxyl is within hydrogen bonding distance of Oδ1 and Oδ2 of Asp-38 and Nη1 of Arg-67 and Arg-73. Finally the endocyclic oxygen of the galactose makes a polar contact with Nη2 of Arg-73. The extensive interactions made by *Ct*CBM55 with the axial *O4* galactose explains why the protein displays tight specificity for this sugar; an equatorial *O4*, evident in glucose, mannose and xylose, would not make polar contacts with the protein and, indeed, is predicted to make steric clashes with the side chain of Asp-38.

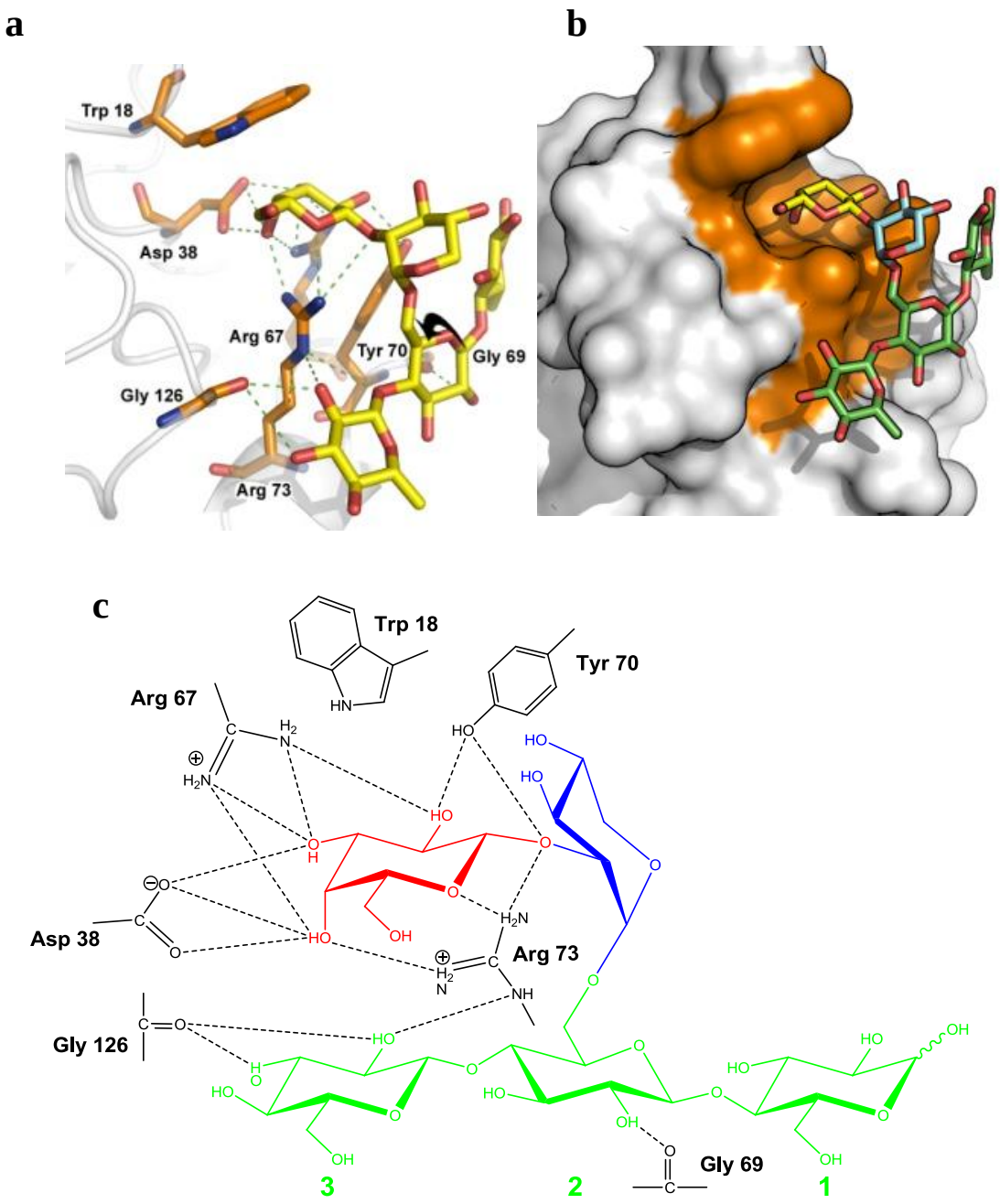


Figure 5.5 Structure of *Ct*CBM55 in complex with xyloglucan
Panel a displays the interactions between *Ct*CBM55 and the ligand, ***Panel b*** provides a surface representation of the ligand binding site occupied by the oligosaccharides, while ***Panel c*** is a schematic representation of the interaction of the carbohydrate molecule with the protein. The sugars are colored as follows: Gal, red; Xyl, blue; Glc, green

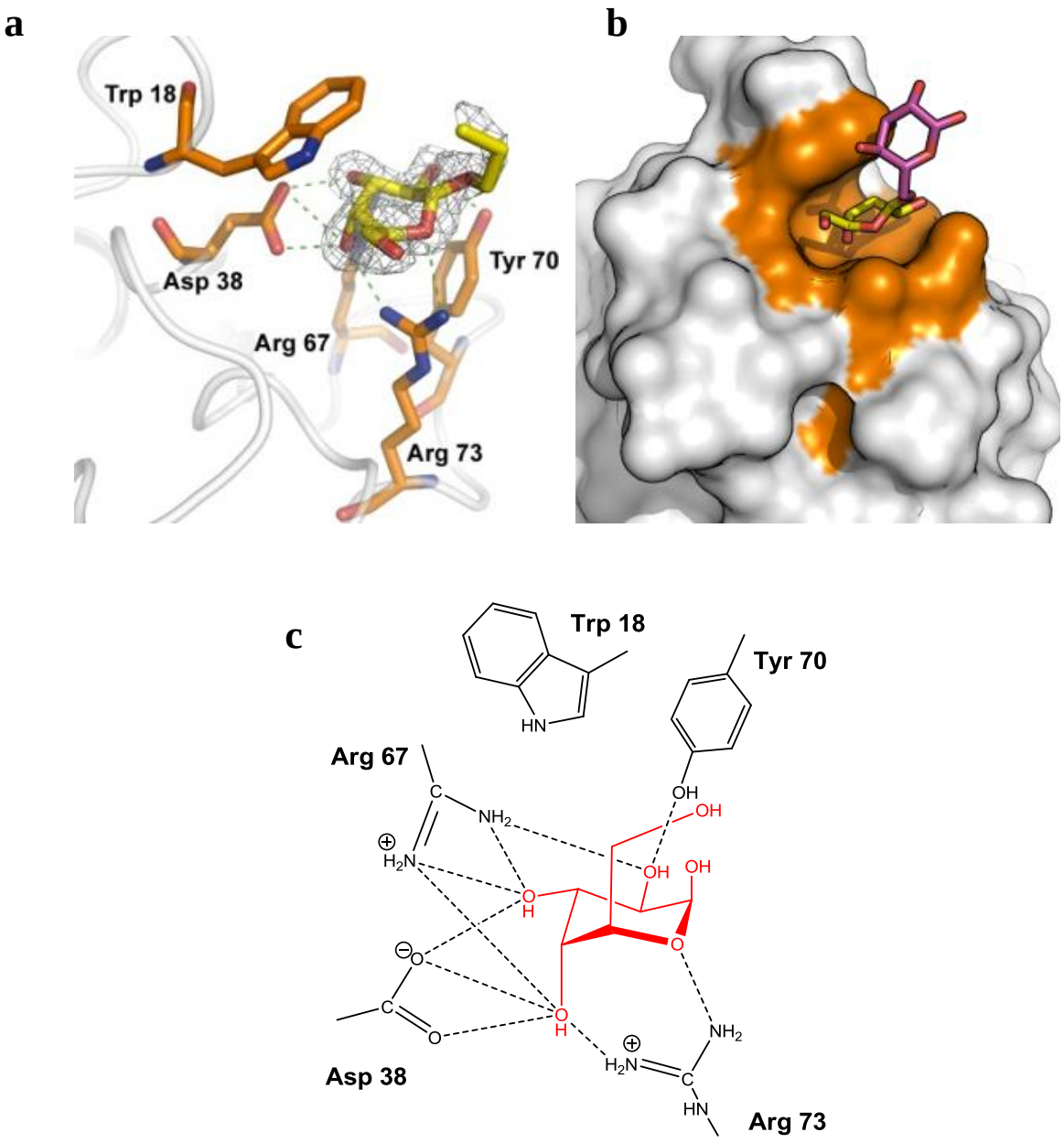


Figure 5.6 Structure of *Ct*CBM55 in complex with galactomannan-derived oligosaccharide (6^1-α-D-$GalMan_3$).
Panel a displays the interactions between *Ct*CBM55 and the ligand, ***Panel b*** provides a surface representation of the ligand binding site occupied by the oligosaccharides, while ***Panel c*** is a schematic representation of the interaction of the carbohydrate molecule with the protein. The sugars are colored as follows: Gal, red; Man is colored red in *Panel c*, yellow in *Panel a.* and magenta in *Panel b.*

Site-directed mutagenesis was deployed to explore the functional significance of the predicted polar and hydrophobic interactions between *Ct*CBM55 and galactose. Replacing with alanine any of the residues that make direct polar contacts (D38A, R67A, Y70A and R73A) or, in the case of Trp-18 (W18A), hydrophobic interactions with galactose, results in the complete loss in binding to xyloglucan, as judged by

affinity gel electrophoresis, Figure 5.7. This loss in ligand binding was confirmed by ITC using xyloglucan as the ligand (data not shown). By contrast, mutation of residues on the concave surface of the protein, W46A, D50A, E53A, Q80A, E85A, F87A and R120A, Figure 5.7, did not appear to cause a reduction in affinity for xyloglucan. Similar observations were made with galactomannan and arabinogalactan (data not shown). These data confirm that the sugar binding site in *Ct*CBM55 is located in the loops connecting the two β-sheets and not on the concave surface of the protein, which is the site of internal polysaccharide recognition in many CBM families.

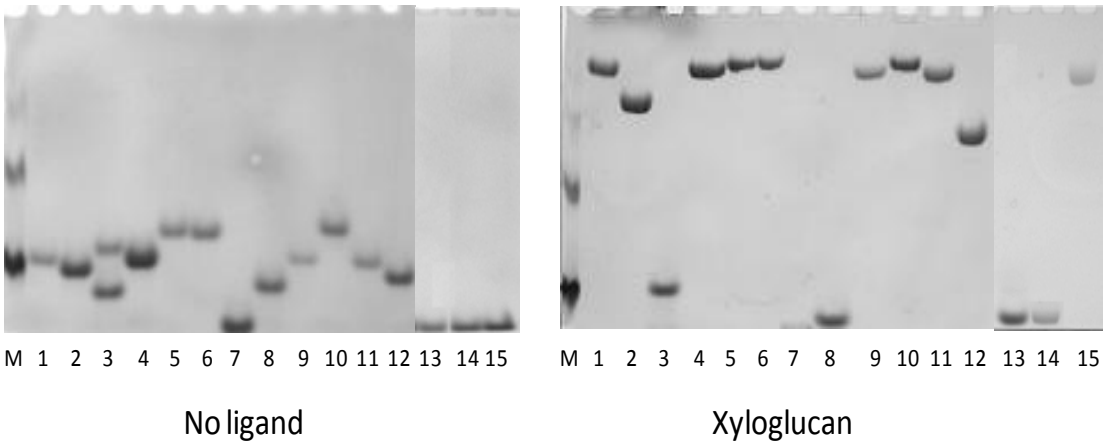


Figure 5.7 Affinity gel electrophoresis of wide type and mutants of *CtCBM55* against xyloglucan. Wide type and mutants of *Ct*CBM52 were electrophoresed on non-denaturating polyacrylamide gels containing no ligand or 1 mg/ml xyloglucan The lanes were as follows: *Ct*CBM52 wide type (1), F36A (2), D38A (3), W46A (4), D50A (5), E53A (6), R67A (7), Y70A (8), Q80A (9), E85A (10), F87A (11), N128A (12), W18A (13), R73A (14) and R120A (15). BSA was used as a non-polysaccharide binding control in lane M.

Significantly *Ct*CBM55 does not make polar contact with the *O6* of galactose suggesting that the protein may interact with L-arabinopyranose in which *O4* is also axial. Data presented in Table 5.4 confirm that *Ct*CBM55 does bind to arabinopyranose, albeit with lower affinity than galactose. This may reflect the strain imposed by forcing the sugar ring into its less preferred pyranose conformation. The crystal structure of *Ct*CBM55 in complex with arabinose reveals that the pentose sugar and galactose make identical interactions with the protein (Figure 5.8).

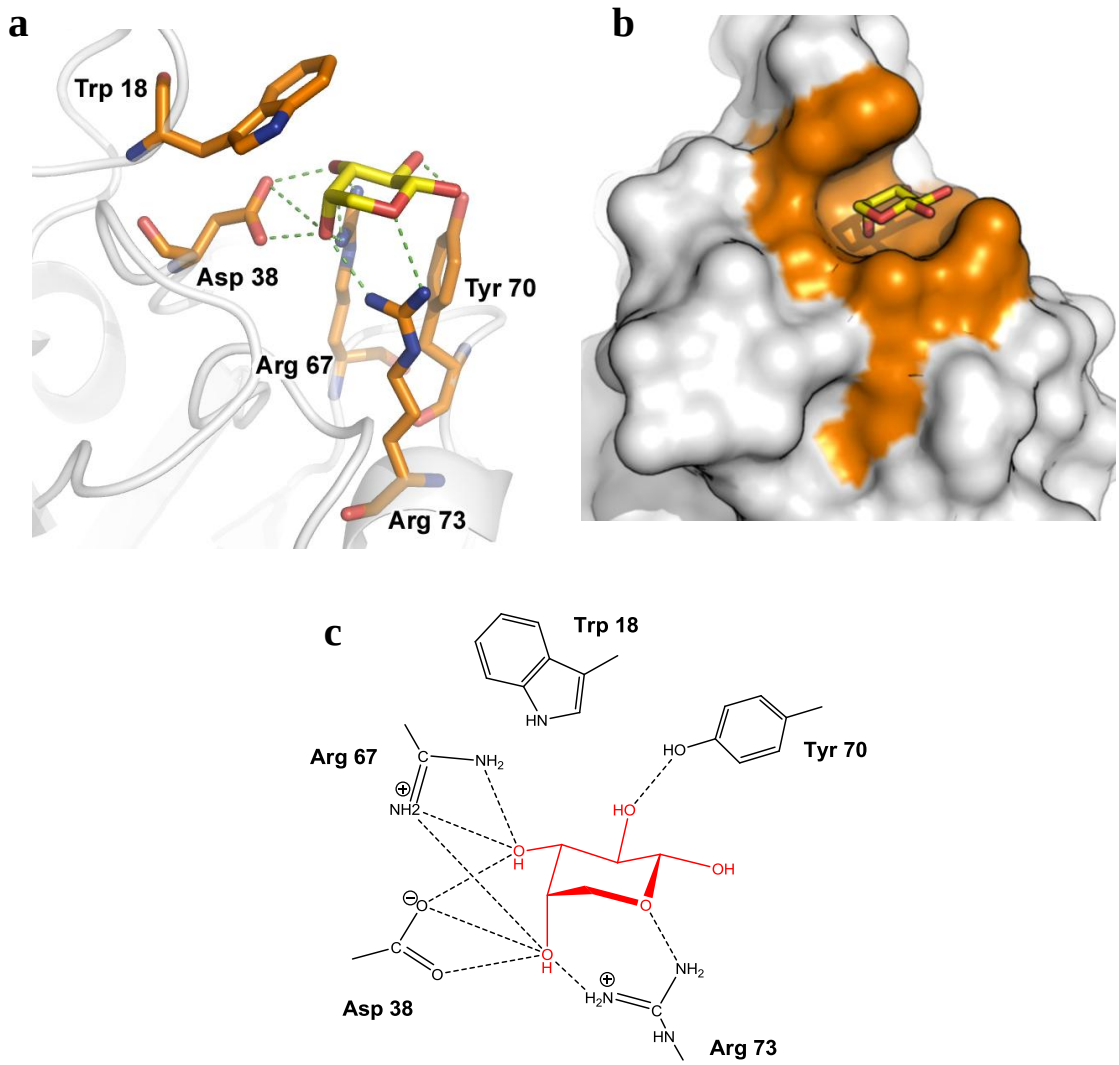


Figure 5.8 Structure of *Ct*CBM55 in complex with arabinopyranose
Panel a displays the interactions between *Ct*CBM55 and L-arabinopyranose, *Panel b* provides a surface representation of the ligand binding site occupied by the sugar, while *Panel c* is a schematic representation of the interaction of the carbohydrate molecule with the protein.

In both ligand complexes, components of the oligosaccharides, in addition to the terminal galactose, are evident. Indeed the successive Gal-βXyl and Xyl-α-Glc linkages curve the xyloglucan oligosaccharide in a shape that is perfectly accommodated by the surface of *Ct*CBM55, which may contribute to the slightly higher affinity of XLXG, compared to galactose, for the protein. The increase in affinity may reflect polar interactions between the Glc backbone and various polar residues on the surface of the protein. To accommodate the Glc or Man backbone of XLXG and $^{1}6$-μ- D-GalMan$_3$, respectively, the side chain of Arg-73 is twisted by 114^{o} which, although imposing an energetic cost, enables both Oε1 and Nε2 to make polar contacts with the backbone of

at least xyloglucan. By contrast the Man backbone of ^{1}6-α- D-GalMan$_3$ did not make direct polar interactions with *Ct*CBM55.

5.3.4. Role of calcium in the function of *Ct*CBM55

Although calcium is not directly implicated in ligand recognition by *Ct*CBM55, several recent studies have shown that the metal ion does contribute to sugar binding in CBMs where the target carbohydrate(s) is accommodated within the loops connecting the β-sheets (Henshaw *et al.*, 2006; Jamal-Talabani *et al.*, 2004; Montanier *et al.*, 2009). To explore the potential role of calcium in the function of *Ct*CBM55, the capacity of the protein to bind to a range of polysaccharides and oligosaccharides in the presence of EDTA (which chelates calcium) and the divalent metal ion, respectively, was assessed. The data, displayed in Table 5.5, show that for simple monovalent ligands, such as galactose and ^{1}6-α-D-GalMan$_3$, neither calcium nor the chelating agent influence carbohydrate recognition. By contrast, in the presence of calcium, *Ct*CBM55 binds to galactose-containing polysaccharides, such as xyloglucan, galactomannan and arabinogalactan >100-fold more tightly than when the metal ion is depleted by EDTA.

Table 5.5 The influence of calcium on the affinity of *Ct*CBM55 for xyloglucan and galactose

Ligand	$K_a \times 10^4$ (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	n[d]
Xyloglucan + calcium[a]	129 ± 14.3	-8.35 ± 0.07	-11.4 ± 0.11	-3.14 ± 0.18	1.00 ± 0.004
Xyloglucan - calcium[b]	0.95 ± 0.02	-5.43 ± 0.01	-11.5 ± 0.15	-6.08 ± 0.16	1.08 ± 0.01
Xyloglucan + EDTA[c]	0.59 ± 0.05	-5.15 ± 0.06	-12.7 ± 1.68	-7.53 ± 1.64	1.04 ± 0.08
Galactose + calcium[a]	0.41 ± 0.003	-3.58 ± 0.01	-9.58 ± 0.12	-6.0 ± 0.18	1.02 ± 0.02
Galactose + EDTA[c]	0.38 ± 0.013	-3.32 ± 0.01	-9.40 ± 0.17	-6.08 ± 0.16	1.12 ± 0.11

[a]*Ct*CBM55 was dialysed in the presence of 2 mM $CaCl_2$ and the metal was added to the ITC experiment.
[b]*Ct*CBM55 dialysed in buffer lacking $CaCl_2$ and ITC was carried out in the absence of calcium.
[c]*Ct*CBM55 dialysed into buffer lacking $CaCl_2$ and the ITC was performed in the presence of 10 mM EDTA.

The data described above indicate that calcium enhances affinity for multivalent ligands but does not influence recognition of carbohydrates containing a single galactose residue. The most likely explanation for the role of calcium is through its capacity to mediate oligomerization of *Ct*CBM55, which would result in enhanced affinity for multivalent ligands through avidity effects. To explore this possibility

*Ct*CBM55 was subjected to size exclusion chromatography in the presence of 5 mM calcium and EDTA, respectively. The Sephadex S100 column used was calibrated with a range of proteins of known molecular weight. In the presence of the chelating agent *Ct*CBM55 migrated as a monomer, however, when EDTA was replaced with calcium the protein appeared to migrate as a dimer. Further support for the calcium-mediated oligomerization of *Ct*CBM55 is provided by the observation that multivalent ligands precipitate when incubated with the protein in the presence of calcium (data not shown), consistent with the capacity of the CBM oligomer to cross link polysaccharide, and even oligosaccharide, chains. Addition of EDTA to these precipitates resulted in their solubilization. To summarize, these data indicate that the capacity of calcium to enhance the affinity of *Ct*CBM55 for multivalent ligands is mediated by metal ion-dependent oligomerization of the protein.

Inspection of the structure of *Ct*CBM55 in the presence of calcium reveals a dimer in the asymmetric unit, Figure 5.9. While there is no evidence that calcium directly cross links the two protomers, the metal ion stabilizes a surface loop that makes polar contacts with the protein partners. Specifically the calcium ion makes coordinate bonds with the O of Lys-27, Oδ1 of Asp-30, O of Asp-32, the backbone O and Oγ of Thr-35, O of Ala-132 and Oε1 of Glu-133, and as such fulfils the octahedral geometry typical of many calcium binding sites. The residues in the loop that comprises the calcium binding site, which interact with the protomer partner, are Gly-31/Leu-33 (the O of Gly-33 makes a hydrogen bond with the N of Leu-33) and Asp-30/Asn-34 (the Nδ2 of Asn-34 makes a polar contact with the backbone carbonyl of Asp-30), Figure 5.9. It should be emphasised that while the two protomers make eight other polar contacts it is likely that the interactions between the two calcium binding sites play a particularly important role in oligomerization. It is formally possible that the protomer interface observed *in crystal* may not fully represent the inter-chain interactions that lead to dimerization in solution. However, the crystal structure of the dimer places the ligand binding sites on opposite faces to the protomer interfaces, which are thus available to interact with different polysaccharide chains, leading to the formation of an insoluble lattice. This is consistent with the distribution of the ligand binding sites on opposite faces of many dimeric lectins, such as human galectin-2 L-14-II, that mediate cross linking of complex carbohydrate targets.

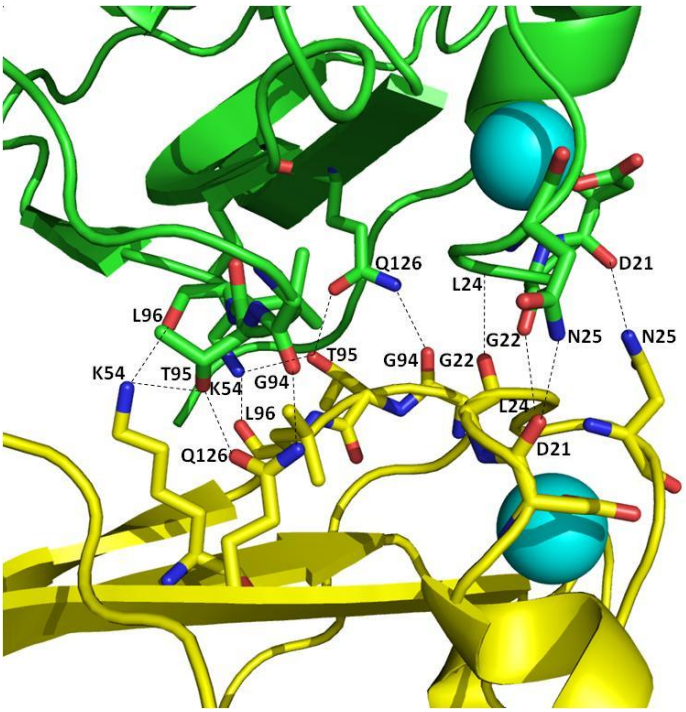


Figure 5.9 The interface between the *Ct*CBM55 protomers in the homodimeric protein
The inter-protomer hydrogen bonds are indicated by a dotted line. Bound calcium is indicated by a cyan sphere.

5.3.5. *Ct*CBM55 defines a new CBM family

*Ct*CBM55 was subjected to BLAST analysis to identify other proteins that are members of this new CBM family. The data, presented in Figure 5.10, reveal 10 protein modules that display >20 % sequence identity with the complete sequence of *Ct*CBM55. These proteins, therefore, are likely to be members of CBM55. The majority of these CBM55 modules are components of large extra-cytoplasmic proteins of unknown function, derived from either mammalian colonic symbionts, such as *Bacteroides fragilis* and *B. thetaiotaomicron*, or the plant cell wall degrading soil bacterium *Chthoniobacter flavus.* The residues that comprise the ligand and calcium binding sites in *Ct*CBM55 are invariant in its three closest homologs, *Chthoniobacter flavus* B9XQP8, *Cytophaga hutchinsonii* Q11SF9 and *Paenibacillus* B1DB96, indicating that these proteins are also highly likely to recognize galactose. Across the CBM55 landscape, however, only two of the five ligand binding residues of *Ct*CBM55 are conserved. These two amino acids, Asp-38 and Arg-73, recognize the axial *O4* of galactose in addition to other components of the ligand. Thus, while it is possible that all the proteins aligned with *Ct*CBM55 recognize galactose, or a derivative of this sugar, it remains to be established whether more distantly related CBM55 members (to *Ct*CBM55) display specificity for this sugar. Ligand recognition in CBM families can

be extremely tight, as in the case of modules that recognize crystalline polysaccharides such as cellulose (Nagy *et al.*, 1998; Raghothama *et al.*, 2000; Tormo *et al.*, 1996), or extremely plastic, exemplified by CBM4 (Boraston *et al.*, 2002) and, more significantly, by CBM6 where even the location of the carbohydrate binding site is variable (see chapter 4 and Czjzek *et al.*, 2001; Henshaw *et al.*, 2006; Henshaw *et al.*, 2004; Tunnicliffe *et al.*, 2005; van Bueren *et al.*, 2005).

Figure 5.10 Sequence alignment of *Ct*CBM55 with related protein modules
BLAST analysis of *Ct*CBM55, corresponding to residues 739 to 885 of Cthe_2193 identified 10 protein modules that displayed >19 % sequence identity, which are identified by their UNIPROT accession number. Residues that are identical in all 10 proteins is indicted by * while those in which the similar amino acid is denoted by :. Residues that were subjected to mutagenesis and belong to the binding cleft are boxed in blue.

5.3.6. Description of the biochemistry and structure of the catalytic module appended to the CBM6 of Cthe_2193

The biochemical properties of the catalytic module of Cthe_2193 was explored with the following complex carbohydrates: barley β-glucan, carboxymethylcellulose, cellohexaose, sugarbeet arabinan, debranched arabinan, potato galactan, lupin galactan, galactomannan, mannan, larchwood arabinogalactan, laminarin, lichenan,

homogalacturonan, rhamnogalacturonan I, glucuronoxylan, oat spelt xylan, wheat and rye arabinoxylan, β1,3-xylan, xylohexaose, xylopentaose, xylotetraose, xyloglucan and xyloglucan oligosaccharide d.p. 14. Assays used 1 % substrate in 50 mM Na/Hepes buffer, pH 7.0, containing 5 mM $CaCl_2$ at 37 ^{o}C. Incubations were for up to 24 h deploying 20 μM enzyme. No detectable reducing sugar could be detected for most of the polysaccharides. However, low levels of reducing sugar were released from xylans, particularly arabinoxylans. HPLC analysis revealed that very low concentrations of a range of xylooligosaccharides were released over 24 h from all the β1,4-xylans, and partial cleavage of xylohexaose and xylopentaose, but not xylotetraose, was evident after prolonged incubations. No small molecular weight products were released from the other polysaccharides and oligosaccharides assessed. The CBM6, designated CBM6-Cthe_2193, and CBM13 of Cthe_2193 were evaluated for ligand binding, in the presence of calcium, using affinity gel electrophoresis and the following polysaccharides: Hydroxyethylcellulose, lichenan, laminarin, crudlan, birchwood xylan, oatspelt xylan, wheat arabinoxylan, rye arabinoxylan, glucuonoxylan, acetylated xylan, ferulic acid xylan, β1,3-xylan, xyloglucan, β-mannan, carob galactomannan, glucomannan, larchwood arabinogalactan, potato galactan, lupin galactan, sugarbeet arabinan, debranched arabinan, rhamnogalacturonan, rhamnogalacturonan I, polygalacturonic acid, pustulan and pullulan. No retardation of the electrophoretic flow of the two proteins was observed for any of the polysaccharides evaluated. The capacity of the protein to bind to complex carbohydrates and monsaccharides was also assessed by isothermal titration calorimetry. The carbohydrates assessed were as follows: oatspelt xylan, xyloglucan, birchwood xylan, rye arabinoxylan, laminarin, xylohexaoase, xylopentaose, cellohexaose, galactose and araninose. The polysaccharides were at 0.5 %, oligosaccharides at 5 mM. The cell containing 33 μM protein in 50 mM Na/Hepes buffer, pH 7.5, containing 5 mM $CaCl_2$. No binding to any of these carbohydrates was evident.

The structure of the catalytic module of Cthe_2193 (designated GH5') fused to the CBM6 was solved as described in Experimental Procedures in the main text. The catalytic module displays limited sequence identity with CAZy enzymes. However, it exhibits greatest structural identity to the GH5 endoglucanase Cel5G from *Pseudoalteromonas haloplanktis* (PDB code 1TVN) with a Z-score of 24.0 and an rmsd of 2.7 Å over ~300 residues and 15 % sequence identity. GH5' displays a $(\alpha/\beta)_8$ fold typical of clan GH-A enzymes, Figure 5.11. A deep and narrow cleft extends across the

surface of the protein on top of the β-barrel, which is likely to be the substrate binding region. Indeed the cleft houses Glu-138 and Glu-246 located at the end of β-strands 4 and 7, respectively, and comprise the catalytic apparatus (these residues are invariant across the whole GH-A landscape and have been shown to be the catalytic acid-base and nucleophile, respectively, in all enzymes evaluated). In addition to Glu-138 and Glu-246, other residues that are conserved in GH5' and CelG, and in a large number of GH5 families, include Asn-137, which makes critical interactions with *O2* of the sugar at the -1 subsite as it moves from its ground state conformation into its transition state. His-220 and Trp-172, which presents a critical hydrophobic platform for the sugar at the -1 subsite, are also highly conserved in clan GH-A.

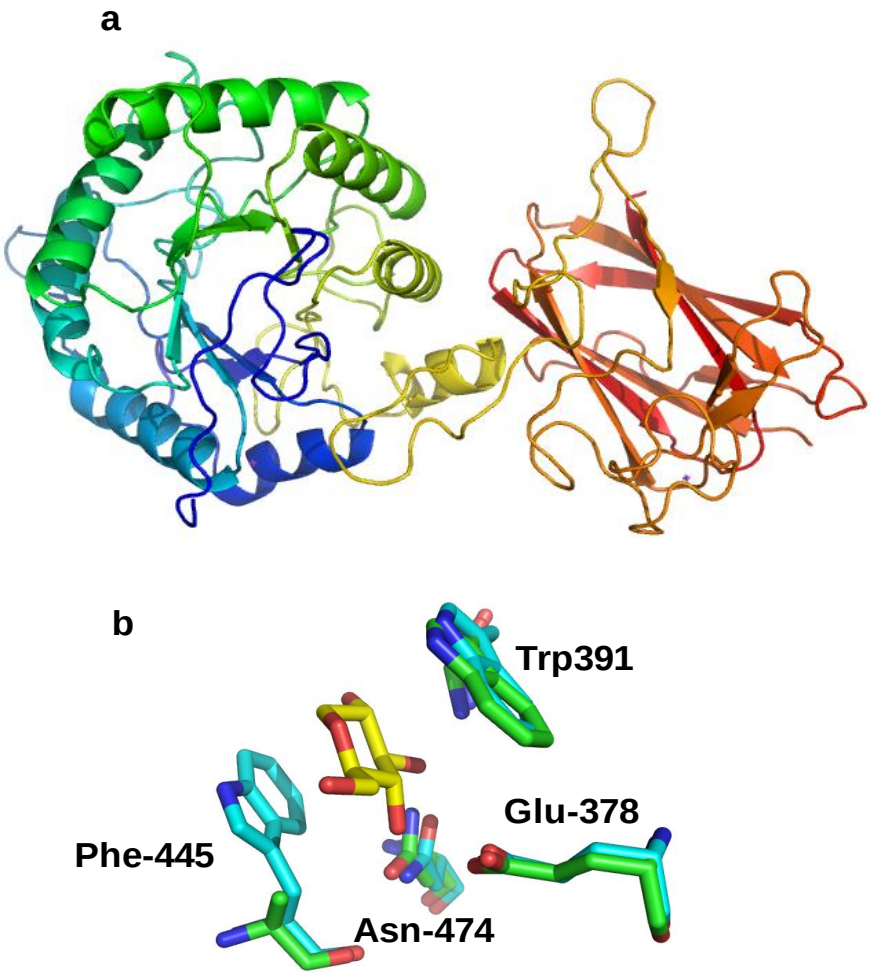


Figure 5.11 The crystal structure of CBM6- Cthe_2193 linked to GH5'
The structure of CBM6- Cthe_2193 linked to GH5' (*Panel a*) is displayed in ribbon representation color-ramped from the **N**-terminus (blue) to the **C**-terminus (red). *Panel b* shows a structural comparison of CBM6-EAM44814 (green) and *Bh*CBM6 (cyan) at the central sugar binding site. The conserved residues corresponding to CBM6-EAM44814 are labeled.

While the structure of GH5' does not reveal the substrate specificity of the enzyme, certain inferences regarding the nature of its substrate and mechanism of action can be deduced. The topology of the active site suggests that the enzyme is an endo-acting glycanase, while its location in clan GH-A demonstrates that the glycosidic bond, which is equatorial to the sugar ring at the -1 subsite (D-β or possibly an L-α sugar), is hydrolyzed by a double displacement or "retaining" mechanism.

CBM6-Cthe_2193 is linked to GH5' by a 36 residue loop (Gly-303-Gly339) that makes extensive contacts with both the catalytic module and the CBM. It is not clear whether this linker sequence contacts the two modules in solution or whether these interactions are a crystallographic artifact. It is evident, however, that the inter-module sequence does not display the repetitive features of known flexible linkers that are typically rich in hydroxyl amino acids or, as occurs frequently in *C. thermocellum* cellulosomal proteins, Pro-Thr repeats. Thus, on balance we believe that the extensive interactions between the linker and the two modules occurs *in vivo*, although the biological significance for such structural rigidity is uncertain.

Typical of family 6 CBMs, CBM6-Cthe_2193 displays a typical β-sandwich fold. Its closest structural homolg is *Bh*CBM6 (PDB code 1w9w), with which is shares 30 % sequence identity and an rmsd of 1.6 Å over 134 residues. *Bh*CBM6 binds to laminarin, a carbohydrate not recognized by CBM6-Cthe_2193. In general the ligand binding site of CBM6s is located in the loops that connect the two β-sheets, and indeed retains many of the features of the proximal region of the ligand binding site of family 6 CBMs. Thus CBM6-Cthe_2193 displays two surface aromatic residues, Trp-391 and Phe-445 capable of aligning parallel to the two faces of the bound sugar (although the aromatic ring of Phe- has not been built its conformation is likely to be locked by the pyranose ring of the bound sugar). At the base of the ligand binding site is Asn-474, which is invariant in CBM6 and makes hydrogen bonds with the sugar ligand located in the centre of the binding site. Indeed the glucose located in the central binding site of CBM6-Cthe_2193 and *Bh*CBM6 will make very similar, if not identical, interactions with their respective proteins, Figure 5.11**.** If this prediction is correct then the CBM6 will bind to a non-reducing terminal sugar (likely glucose or xylose) as the ligand binding site displays a blind canyon topology, with *O3* and *O4* pointing at the protein surface.

5.4. DISCUSSION

This report reveals a novel CBM family which recognizes galactose-containing polysaccharides, located in a protein that is a component of the *C. thermocellum* cellulosome. The β-sandwich fold displayed by *Ct*CBM55 is typical of the majority of CBM families (Boraston *et al.*, 2004). While, historically, the ligand binding site of CBMs is associated with the concave surface of these proteins, this appears to apply only to proteins that recognize the internal regions of polysaccharides. The location of the *Ct*CBM55 ligand binding site in the loops that connect the β-sheets is an example of an increasingly common feature of CBMs that recognize terminal sugars. The tight specificity for galactose is mediated by numerous interactions with the axial *O4*, which distinguishes the sugar from other neutral carbohydrates, notable mannose, xylose and glucose, which are widely distributed in plant structural polysaccharides (Brett *et al.*, 1996).

The observation that *Ct*CBM55 binds more tightly to complex polysaccharides, than to monovalent ligands, and that binding is associated with the formation of an insoluble polysaccharide lattice, is a classic feature of avidity effects (Boraston *et al.*, 2002; Vijayan *et al.*, 1999). The rationale for the increase in K_a relates to the decrease in the frequency with which the protein will completely dissociate from its ligand complex, compared to monovalent carbohydrates (Charnock *et al.*, 2002; Freelove *et al.*, 200). Examples of such avidity effects have been observed previously in enzymes containing multiple CBMs that recognize the same complex ligand. Thus multiple copies of xylan binding CBM6s and CBM2bs, present in several bacterial xylanases, bind more tightly to their target polysaccharide relative to the individual modules (Bolam *et al.*, 2001; Boraston *et al.*, 2002). By contrast avidity effects through the oligomerization of a discrete enzyme-located CBM, as observed here for *Ct*CBM55, have not been reported previously. It should be noted, however, that CBM29-2, derived from the non-catalytic *Piromyces equi* protein, NCP1, binds more tightly to polysaccharides than oligosaccharides that occupy the complete binding site (Charnock *et al.*, 2002). While this differential binding indicates that the protein is able to oligomerize, the biological relevance of this observation is unclear as it may simply reflect the capacity of CBM29-2 to interact with its structural and biochemical homolog, CBM29-1, which is also a component of NCP1 (Freelove *et al.*, 2001).

The observation that *Ct*CBM55 requires calcium for tight binding to polysaccharides is consistent with several recent reports showing the importance of the

metal ion in ligand recognition. Thus the xylan specificity displayed by a CBM36 from *Paenbacillus polymyxa* xylanase 43A is calcium dependent (Jamal-Talabani *et al.*, 2004), while a group of CBM35s only bind to uronic acids in the presence of the metal ion (Montanier *et al.*, 2009). The crystal structures of the CBM35 and CBM36 modules in complex with their respective ligands show that calcium makes direct interactions with the bound carbohydrate. In contrast, while the CBM6 module Aga16B-CBM6-2 also requires calcium to bind to its ligand, neoagarobiose, the metal does not bind directly to the disaccharide but stabilizes a loop that contain residues which make polar and hydrophobic contacts with the target carbohydrate (Henshaw *et al.*, 2006). Here *Ct*CBM55 reveals a further mechanism by which calcium contributes to CBM ligand recognition. The metal ion mediates tight binding to polysaccharides through its indirect contribution to protein oligomerization (stabilization of a loop at the protomer interface), enabling a CBM dimer to bind to multiple galactose residues in multivalent sites leading to avidity effects. CBM-mediated oligomerization of enzymes has not been previously reported and, indeed, all known glycoside hydrolases that contain CBMs are monomeric.

In conclusion, this report reveals a novel family of CBMs, the founding member of which is a component of the *C. thermocellum* cellulosome. This multienzyme complex is anchored onto the cell wall through CBM3a that binds extremely tightly to crystalline cellulose and, because it is a component of CipA, is present in all cellulosome molecules. The catalytic subunits appended to CipA provide a further level of specificity by recruiting these macromolecules, through their appended CBMs, to specific regions of the plant cell wall. For example, the xylanases *Ct*Xyn10A and *Ct*Xyn11A contain xylan-specific CBMs and thus target the cellulosome towards regions rich in this xylose polymer, while pectin modifying enzymes and xyloglucanses contain CBMs that direct these enzymes to their target substrates. This report reveals a novel CBM family that shows specificity for terminal D-galactose and L-arabinopyranose residues, specificity not previously observed in CBMs that are components of the *C. thermocellum* cellulosome. The CBM, uniquely, mediates protein dimerization, which confers polysaccharide, rather than oligosaccharide, targeting through avidity effects. The implication of this CBM-mediated protein recruitment, in relation to the activity of the appended catalytic module, is currently unclear.

CHAPTER 6 General Discussion and Future Perspectives

Plant cell walls are composed of an intricate network of polysaccharides, joined together by several molecular forces. Its degradation is one of the most important mechanisms for achieving carbon recycling in nature. Carbohydrate esterases (CEs), Glycoside hydrolases (GHs) and pectate lyases (PLs), as well as their associated carbohydrate-binding modules (CBMs), are involved in plant cell wall degradation and this work is aimed to elucidate some unclear issues concerning their mechanisms of action.

The majority of the matrix plant cell wall polysaccharides, including xylan, mannan, glucomannan and pectin, can be acetylated which sterically restricts the access of the repertoire of glycoside hydrolases that attack the glycosidic linkages in these polymers (Brett *et al.*, 1996). Removal of the acetyl groups though the action of carbohydrate esterases greatly increases the efficiency of polysaccharide saccharification through the action of glycoside hydrolases (Ferreira *et al.*, 1993). Thus, acetyl esterases are important components of the plant cell wall-degrading apparatus of saprophytic microorganisms. Data presented in chapter 2 demonstrates that *Ct*Ces3-1 is an acetyl esterase that displays a high level of specificity for acetylated xylan and xylooligosaccharides and this result is entirely consistent with the other CE3 esterases characterized to date, which also hydrolyse acetylated xylan (Dalrymple *et al*, 1997; Aurilia *et al.*, 2000). Thus, the likely biological function of *Ct*Ces3 is to increase access of the endo-acting xylanases for the undecorated xylan backbone and this may be particularly important as extensive deacetylation of xylans, which is not associated with hydrolysis of the xylan backbone, will result in the precipitation of the polysaccharide. An intriguing feature of *Ct*Ces3 is that the enzyme contains two esterase modules that display 97% of sequence identity and, by inference, will display the same substrate specificity. So, if two catalytic modules can attack the same xylan chain, an avidity effect will occur (Freelove *et al.*, 2001; Boraston, *et al.*, 2002), and thus the enzyme will cleave more effectively the polysaccharide side chain. This cooperativity is likely to confer specificity for xylan rather than xylooligosaccharides.

The biological function of CBMs from a variety of proteins was explored in the last three chapters. In general, there is a tight correlation between the ligands recognized by the CBMs and the substrate specificity of the appended catalytic modules. Thus CBMs, by recruiting their appended enzymes onto the surface of the target substrate, potentiate catalysis particularly against insoluble substrates. In Chapter 3, the architecture of the ligand binding sites of four distinct family 35 CBMs, appended to three plant cell wall hydrolases and an enzyme, CsxA, which confers protection from an antibacterial fungal polysaccharide was shown to be remarkably similar. The four domains display similar binding specificity for glucuronic acid and/or Δ4,5-anhydrogalaturonic acid (Δ4,5-GalA). It was shown that despite displaying significant structural and biochemical similarities, there is considerable diversity in the biological functions of the four CBM35s. Although three of the CBM35s target the plant cell wall, notably Δ4,5-GalA that is actively degraded, the capacity of Chi-CBM35 to function as a bacterial cell wall attachment module is perhaps the most striking feature of this protein ensemble. Proteins can be retained on the surface of eukaryotic and prokaryotic cells through a variety of mechanisms that do not feature carbohydrate recognition. Thus, the observation that Chi-CBM35 tethers CsxA to the surface of *A. orientalis* by binding to a carbohydrate ligand is an unusual biological role for a CBM and carbohydrate-binding proteins in general. The target ligand for Chi-CBM35 in the cell wall of *A. orientalis* is most likely an uronic acid, such as GlcA. Although the ITC-derived affinity of Chi- CBM35 for the uronic acid (KD ≈ 100 μM) is modest, this solution affinity does not reflect the binding strength on a membrane surface where clustering of the ligands can result in a marked increase in apparent affinity. The biological rationale for this function appears to be to keep the enzyme in proximity to the bacterium and by presenting the enzyme on the surface of the bacterium, only the cationic fragments of chitosan, which pose the most immediate threat to the bacterium, are degraded by CsxA. The capacity of the other 3 CBM35s, which are components of plant cell wall degrading enzymes, to bind Δ4,5-GalA indicates that they play a role in targeting the hydrolases and esterases to regions of their composite substrate that are accessible to biological degradation. It should be emphasized, however, that Xyl-CBM35 also displays significant affinity for GlcA, a side chain present in several xylans (Pena *et al.,* 2007), and thus this plasticity in uronic acid recognition provides the enzyme with a flexible targeting strategy that is tailored to the plant biomass presented to the bacterium. It is possible that Xyl-CBM35 initially directs the xylanase toward

regions of cell walls that are being actively degraded but, as xylan structures are revealed, the enzyme is shuttled onto the hemicellulosic polysaccharide affording the enzyme access to its target substrate. Data presented in Chapter 3 demonstrates how nature has exploited convergent binding specificity for divergent function. It is evident, therefore, that the context of the ligands recognized by CBMs is central to their biological function. In addition to providing insights into the biological role of CBMs as enzyme delivery molecules and host cell adhesins, respectively, data presented here also highlights the potential exploitation of these protein modules within an industrial setting. The novel specificities displayed by the family 35 CBMs will not only contribute to the toolbox of biocatalysts required to generate biofuels from lignocellulose, but also represent valuable probes that can be deployed in the analysis of plant cell wall architecture.

Cellvibrio mixtus is a saprophytic mesophilic bacterium that colonizes the soil and efficiently degrades a variety of plant cell wall polysaccharides. An enzyme derived from this bacterium, *Cm*Cel5B, presents a modular arquitecture containing an N-terminal family 5 GH catalytic module, followed by an internal CBM6, termed CBM6-1, and a C-terminal CBM6, termed CBM6-2. In chapter 4, CBM6-1 from *Cm*Cel5B was shown to bind β1,3-glucans through residues located both in clefts A and B. In addition, the ligand binding interface is suggested to accommodate more than six β1,3-linked glucose residues. The biological significance of this specificity is intriguing since β1,3-glucans are not present in the cell walls of terrestrial plants (Cid *et al.*, 1995). However, it is possible that *C. mixtus* hydrolyses fungal cell wall polysaccharides and as fungi are the initial colonizers of plant biomass, it is possible that saprophytic bacteria could have developed fungal lytic mechanisms allowing them to compete with fungal species that also use plant cell wall carbohydrates as an energy source. Intriguingly, the catalytic module appended to CBM6-1 consists of a typical cellulase arguing against this possibility. Therefore, CBM6-1 may target *Cm*Cel5B to plant cell wall locations that are being actively degraded by fungi where a variety of complex carbohydrates are accessible to enzyme attack. When the enzyme is located in the vicinity of the fungal cell wall, the action of the microbial eukaryote may expose plant cell wall β1,3-1,4-glucans, which are targeted by the C-terminal CBM6-2 of the glucanase, enabling the catalytic module to degrade its target substrate. Therefore, the cooperative action of CBM6-1 and CBM6-2 allows *Cm*Cel5B to be directed to the region of the cell wall that

are actively being degraded by fungi, through CBM6-1, where the catalytic module, in harness with CBM6-2, will hydrolyse plant cell wall β1,3-1,4-glucans. Finally, it is also possible that the primary substrates targeted by *Cm*Cel5B are the β1,3-1,4-glucans located in the fungal cell wall itself. Chitin, galactomannan and the linear β1,3-1,4-glucans were found to be covalently linked to the non-reducing end of β1,3-glucans, explaining the specificity of CBM6-1 (Fontaine *et al.*, 2002). The prevalence of β1,3-1,4-glucans in the cell wall of cellulolytic and hemicellulolytic fungi, however, is currently unknown.

Chapter 5 reveals a novel CBM family which recognizes galactose-containing polysaccharides, located in a protein that is a component of the *C. thermocellum* cellulosome. The β-sandwich fold displayed by *Ct*CBM55 is typical of the majority of CBM families (Boraston *et al.*, 2004). The location of the *Ct*CBM55 ligand binding site in the loops that connect the β-sheets is an example of an increasingly common feature of CBMs that recognize terminal sugars. The tight specificity for galactose is mediated by numerous interactions with the axial *O4*, which distinguishes the sugar from other neutral carbohydrates, notably mannose, xylose and glucose, which are widely distributed in plant structural polysaccharides (Brett *et al.*, 1996). We also observed that *Ct*CBM55 binds more tightly to complex polysaccharides than to monovalent ligands, and that binding is associated with the formation of an insoluble polysaccharide lattice, which is a classic feature of avidity effects (Boraston *et al.*, 2002; Vijayan *et al.*, 1999). Examples of such avidity effects have been observed previously in enzymes containing multiple CBMs that recognize the same complex ligand. Thus multiple copies of xylan binding CBM6s and CBM2bs, present in several bacterial xylanases, bind more tightly to their target polysaccharide relative to the individual modules (Bolam *et al.*, 2001; Boraston *et al.*, 2002). In contrast, avidity effects through the oligomerization of a discrete enzyme-located CBM, as observed here for *Ct*CBM55, have not been reported previously.

Although calcium is not directly implicated in ligand recognition by *Ct*CBM55, several recent studies have shown that the metal ion does contribute to sugar binding in CBMs, where the target carbohydrate(s) is accommodated within the loops connecting the β-sheets. Therefore, the observation that *Ct*CBM55 requires calcium for tight binding to polysaccharides is consistent with this recent reports showing the importance of the metal ion in ligand recognition. *Ct*CBM55 revealed a further mechanism by

which calcium contributes to CBM ligand recognition. The metal ion mediates tight binding to polysaccharides through its indirect contribution to protein oligomerization (stabilization of a loop at the protomer interface), enabling a CBM dimer to bind to multiple galactose residues in multivalent ligands leading to avidity effects. CBM-mediated oligomerization of enzymes has not been previously reported and, indeed, all known GHs that contain CBMs are monomeric. In conclusion, Chapter 5 reveals a novel family of CBMs, the founding member of which is a component of the *C. thermocellum* cellulosome. *Ct*CBM55 display specificity for terminal D-galactose and L-arabinopyranose residues a specificity not previously observed in CBMs that are components of the *C. thermocellum* cellulosome. The CBM, uniquely, mediates protein dimerization, which confers polysaccharide, rather than oligosaccharide, targeting through avidity effects. The implication of this CBM-mediated protein recruitment, in relation to the activity of the appended catalytic module, is currently unclear.

References

Abbott, D. W. and Boraston, A.B. (2007) Specific recognition of saturated and 4,5-unsaturated hexuronate sugars by a periplasmic binding protein involved in pectin catabolism. *J Mol Biol* 369:759–770.

Ali, M. K., Hayashi, H., Karita, S., Goto, M., Kimura, T., Sakka, K. and Ohmiya, K. (2001) Importance of the carbohydrate-binding module of Clostridium stercorarium Xyn10B to xylan hydrolysis. *Biosci. Biotechnol. Biochem.* 65, 41–47

Ali, E., Araki, R., Zhao, G., Sakka, M., Karita, S., Kimura, T. and Sakka, K. (2005) Functions of family-22 carbohydrate-binding modules in Clostridium josui Xyn10A. *Bioscience, Biotechnology, and Biochemistry*. 69, 6

Araki, R., Ali, M. K., Sakka, M., Kimura, T., Sakka, K. and Ohmiya, K. (2004) Essential role of the family-22 carbohydrate-binding modules for β-1,3→1,4-glucanase activity of Clostridium stercorarium Xyn10B. *FEBS Lett.* 561, 155–158

Aurilia, V., Martin, J. C., McCrae, S. I., Scott, K. P., Rincon, M. T. and Flint, H. J. (2000) Three multidomain esterases from the cellulolytic rumen anaerobe Ruminococcus flavefaciens 17 that carry divergent dockerin sequences. *Microbiology*, 146, 1391–1397.

Bartnicki-Garcia, S. (1968) Cell wall chemistry, morphogenesis, and taxonomy of fungi. *Annu Rev Microbiol* 22: 87-108

Bayer, E. A., Kenig, R. and Lamed, R. (1983) Adherence of Clostridium thermocellum to cellulose. *J. Bacteriol*. 156, 818–827.

Bayer, E. A., Belaich, J. P., Shoham, Y. and Lamed, R. (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol*. 58, 521–554.

Bayer, E. A., Lamed, R., White, B. A. and Flint, H. J. (2008) From cellulosomes to cellulosomics. Chemical record, 8, 364-377

Blair, D. E., Schuttelkopf, A. W., MacRae, J, A., van Aalten, D. M. F. (2005) Structure and metal-dependent mechanism of peptidoglycan deacetylase, a Streptococcal virulence factor. *Proc Natl Acad Sci USA*, 102, 15429 - 15434.

Blair, D. E., Hekmat, O., Schuttelkopf, A.W., Shrestha, B., Tokuyasu, K., Withers, S. G. & van Aalten, D. M. (2006) Structure and mechanism of chitin deacetylase from the fungal pathogen Colletotrichum lindemuthianum. *Biochemistry*, 45, 9416–9426.

Black, G. W., Rixon, J. E., Clarke, J. H., Hazlewood, G. P., Ferreira, L. M., Bolam, D. N. and Gilbert, H. J. (1997) Cellulose binding domains and linker sequences potentiate the activity of hemicellulases against complex substrates. *J. Biotechnol.* 57, 59-69

Bolam, D. N., Ciruela, A., Mcqueen-Mason, S., Simpson, P., Williamson, M. P., Rixon, J. E., Boraston, A., Hazlewood, G. P. and Gilbert, H. J. (1998) Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochem. J.* 331, 775–781

Bolam, D. N., Xie, H., White, P., Simpson, P. J., Hancock, S. M., Williamson, M. P. and Gilbert, H. J. (2001) Evidence for synergy between family 2b carbohydrate binding modules in Cellulomonas fimi xylanase 11A. *Biochemistry*. 40(8), 2468-2477.

Bolam, D. N., Xie, H., Pell, G., Hogg, D., Galbraith, G., Henrissat, B., and Gilbert H.J. (2004) X4 modules represent a new family of carbohydrate-binding modules that display novel properties. *J. Biol. Chem.* 279, 22953-22963

Boraston, A. B., Tomme, P. et al. (2000) A novel mechanism of xylan binding by a lectin-like module from Streptomyces lividans xylanase 10A. *Biochem J* 350 Pt 3: 933-41.

Boraston, A. B.,McLean, B.W.,Chen,G., Li,A.,Warren, R. A. and Kilburn, D. G. (2002a) Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase. *Mol. Microbiol.* 43, 187–194.

Boraston, A. B., Nurizzo, D., Notenboom, V., Ducros, V., Rose, D. R., Kilburn, D. G. and Davies, G. J. (2002b) Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *J Mol Biol.* 319, 1143-1156

Boraston, A. B., Kwan, E., Chiu, P., Warren, R. A. J. and Kilburn, D. G. (2003a) Recognition and hydrolysis of noncrystalline cellulose. *J. Biol. Chem.* 278, 6120–6127

Boraston, A. B., Notenboom, V., Warren, R. A., Kilburn, D. G., Rose, D. R. and Davies, G. (2003b) Structure and ligand binding of carbohydrate-binding module *Cs*CBM6-3 reveals similarities with fucose-specific lectins and "galactose-binding" domains. *J. Mol. Biol.* 327, 659-669.

Boraston, A. B., Revett, T. J., Boraston, C. M., Nurizzo, D., and Davies, G. J. (2003c) Structural and thermodynamic dissection of specific mannan recognition by a carbohydrate binding module, TmCBM27. *Structure* 11, 665-675

Boraston, A. B., Bolam, D.N., Gilbert, H. J., Davies, G. J. (2004) Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. *Biochem J* 382:769–781.

Boraston, A. B., Wang, D., Burke, R. D. (2006) Blood group antigen recognition by a *Streptococcus pneumoniae* virulence factor. *J Biol Chem* 281:35263–35271.

Boudet, A. M., Kajita, S., Grima-Pettenati, J. and Goffner, D. (2003) Lignins and lignocellulosics: a better control of synthesis for new and improved uses. *Trends Plant Sci.* 8, 576-581.

Brett, C. T. and Waldren, K. (1996) Physiology and Biochemistry of Plant Cell Walls. Topics in Plant Functional Biology. (Black, M. & Charlewood, B., eds), vol. 1 Chapman and Hall, London.

Brown, I. E., Mallen, M. H., Charnock, S. J., Davies, G. J. and Black, G. W. (2001) Pectate lyase 10A from Pseudomonas cellulosa is a modular enzyme containing a family 2a carbohydrate-binding module. *Biochemical Journal.* 355, 11

Brunger AT (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355:472–475.

Bueren, A. L., Morland, C., Gilbert, H. J., and Boraston, A. B. (2005) Family 6 carbohydrate binding modules recognize the non-reducing end of beta-1,3-linked glucans by presenting a unique ligand binding surface. *J. Biol. Chem.* 280, 530-537

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37, D233-238

Carvalho, A. L., Dias, F. M. V., Prates, J. A. M., Nagy, T., Gilbert, H. J., Davies, G. J., Ferreira, L. M. A., Romao, M. J. and Fontes, C. M. G. A. (2003) Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13809-13814.

Carvalho, A. L., Goyal, A., Prates, J. A. M., Bolam, D. N., Gilbert, H. J., Pires, V. M. R., Ferreira, L. M. A., Romão, M. J. and Fontes, C. M. G. A (2004) Crystal structure and functional properties of the family 11 Carbohydrate-Binding Module of cellulosomal cellulase Lic26A-Cel5E of *Clostridium thermocellum*. *J. Biol. Chem*. 279, 34785-34793.

Cavanagh JB (1999) Corpora-amylacea and the family of polyglucosan diseases. *Brain Res Rev* 29:265–295.

Chatterjee, D., Aspinall, G. O., Brennan, P. J. (1987) The presence of novel glucuronic acidcontaining, type-specific glycolipid antigens within Mycobacterium spp. Revision of earlier structures. *J Biol Chem* 262, 3528–3533.

Chatterjee, D., Bozic, C., Aspinall, G. O., Brennan, P. J. (1988) Glucuronic acid- and branched sugar-containing glycolipid antigens of *Mycobacterium avium*. *J Biol Chem* 263, 4092– 4097.

Charnock, S. J., Lakey, J. H., Virden, R., Hughes, N., Sinnott, M. L., Hazlewood, G. P., Pickersgill, R. and Gilbert, H. J. (1997) Key residues in subsite F play a critical role in the activity of Pseudomonas fluorescens subspecies cellulosa xylanase A against xylooligosaccharides but not against highly polymeric substrates such as xylan. *J Biol Chem.* 272, 2942-2951

Charnock, S. J., Bolam, D. N., Turkenburg, J. P., Gilbert, H. J., Ferreira, L. M., Davies, G. J. and Fontes, C. M. (2000) The X6 "thermostabilizing" domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the Clostridium thermocellum X6b domain. *Biochemistry* 39, 5013–5021

Charnock, S. J., Bolam, D. N., Nurizzo, D., Szabo, L., McKie, V. A., Gilbert, H. J. and Davies, G. J. (2002) Promiscuity in ligand-binding: The three-dimensional structure of a *Piromyces* carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proc. Natl. Acad. Sci. U S A.* 99, 14077-14082

Centeno, M. S., Goyal, A., Prates, J. A., Ferreira, L. M., Gilbert, H. J. and Fontes, C. M. (2006) Novel modular enzymes encoded by a cellulase gene cluster in Cellvibrio mixtus. *FEMS Microbiol Lett* 265, 26-34.

Cid, V. J., Duran, A., Del Rey, F., Snyder, M. P., Nombela, C. and Sanchez, M. (1995) Molecular basis of cell integrity and morphogenesis in *Saccharomyces cerevisiae*. *Microbiol. reviews* 59, 345-386.

Coggins BE, Li X, McClerren AL, Hindsgaul O, Raetz CRH, Zhou P (2003) Structure of the lpxC deacetylase with a bound substrate analog inhibitor. *Nat Struct Biol*, 10, 645-651.

Collaborative Computational Project Number 4′ (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D, 50, 760–763.

Cosgrove, D. J. (2005) Growth of the plant cell wall. *Nat Rev Mol Cell Biol.* 6(11): 850-61.

Cote, N., A. Fleury, et al. (2006) Two exo-beta-D-glucosaminidases/exochitosanases from actinomycetes define a new subfamily within family 2 of glycoside hydrolases. *Biochem J* 394(Pt 3), 675-86.

Coutinho, P. M. and Henrissat, B. (1999) The modular structure of carbohydrate-binding enzymes: an integrated database approach. Recent Advances in Carbohydrate Bioengineering (Gilbert, H. H., Davies, G., Henrissat, B., and Svensson, B., eds), pp. 3-12, The Royal Society of Chemistry, Cambridge, UK

Cowtan, K. (1994) ‘dm’: an automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, 31, 34–38.

Creagh, A. L., E. Ong, et al. (1996) Binding of the cellulose-binding domain of exoglucanase Cex from Cellulomonas fimi to insoluble microcrystalline cellulose is entropically driven. *Proc. Natl. Acad. Sci. USA* 93 (22), 12229-12234

Czjzek, M., Bolam, D. N., Mosbah, A., Allouch, J., Fontes, C. M. G. A. , Ferreira, L. M. A., Bornet, O., Zamboni, V., Darbon, H., Smith, N., Black, G. W., Henrissat, B. and Gilbert, H. J. (2001) The location of the ligand binding site of carbohydrate binding modules that have evolved from a common sequence is not conserved. *J. Biol. Chem.* 276, 48580-48587.

Dalrymple, B. P., Cybinski, D. H., Layton, I.,McSweeney, C. S., Xue, G. P., Swadling, Y. J. & Lowry, J. B. (1997) Three Neocallimastix patriciarum esterases associated with the degradation of complex polysaccharides are members of a new family of hydrolases. *Microbiology*, 143, 2605–2614.

Davies, G. J., T. M. Gloster, et al. (2005) Recent structural insights into the expanding world of carbohydrate-active enzymes. *Curr. Opin. Struct. Biol.* 15(6): 637-45.

Dias, F. M. V., Vincent, F., Pell, G., Prates, J. A. M., Centeno, M. S.J., Tailford, L. E, Ferreira, L. M. A., Fontes, C. M. G. A., Davies, G. J. and Gilber, H. J. (2004) Insights into the molecular determinants of substrate specificity in glycoside hydrolase family 5 revealed by the crystal structure and kinetics of *Cellvibrio mixtus* mannosidase 5A. *J. Biol. Chem.* 279, 25517-25526.

Ding, S. Y., Bayer, E. A., Steiner, D., Shoham, Y. and Lamed, R. (1999) A novel cellulosomal scaffoldin from Acetivibrio cellulolyticus that contains a family 9 glycosyl hydrolase. *J Bacteriol* 181(21): 6720-9.

Ding, S. Y., Rincon, M. T., Lamed, R., Martin, J. C., McCrae, S. I., Aurilia, V., Shoham, Y., Bayer, E. A. and Flint, H. J. (2001) Cellulosomal scaffoldin-like proteins from Ruminococcus flavefaciens. *J Bacteriol* 183(6): 1945-53.

Doi, R. H., Kosugi, A., Murashima, K., Tamaru, Y. and Han, S. O. (2003) Cellulosomes from mesophilic bacteria. *J Bacteriol* 185(20): 5907-14.

Dvortsov, I. A., Lunina, N. A., Chekanovskaya, L. A., Schwarz, W. H., Zverlov, V. V. and Velikodvorskaya, G. A. (2009) Carbohydrate binding properties of a separately folding protein module from {beta}-1,3-glucanase Lic16A of Clostridium thermocellum. *Microbiology* (Print)

Emsley, P. and Cowtan, K. (2004) "Coot: model-building tools for molecular graphics." *Acta Crystallogr. Biol. D*, 60, 2126–2132.

Esnouf, R. M. (1997) An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J Mol Graphics Model* 15, 132–134.

Evans, P. (1993) Data reduction. In CCP4 Daresbury Study Weekend: Data Collection and Processing, DL/SCI/ R34, pp. 114–122, Central Laboratory of the Research Councils, Daresbury Laboratory, Warrington, UK.

Evans, G., Bricogne, G. (2002) Triiodide derivatization and combinatorial counter-ion replacement: Two methods for enhancing phasing signal using laboratory Cu K alpha X-ray equipment. *Acta Crystallogr D* 58, 976–991.

Ezer, A., Matalon, E., Jindou, S., Borovok, I., Atamna, N., Yu, Z., Morrison, M., Bayer, E. A. and Lamed, R. (2008) Cell surface enzyme attachment is mediated by family 37 carbohydrate-binding modules, unique to Ruminococcus albus. *J Bacteriol* 190(24), 8220-2.

Fernandes, A. C., Fontes, C. M., Gilbert, H. J., Hazlewood, G. P., Fernandes, T. H. and Ferreira, L. M. (1999) Homologous xylanases from Clostridium thermocellum: evidence for bi-functional activity, synergism between xylanase catalytic modules and the presence of xylan-binding domains in enzyme complexes. *Biochem. J.* 342, 105-110.

Ferreira, L. M., Durrant, A. J., Hall, J., Hazlewood, G. P. and Gilbert, H. J. (1990) Spatial separation of protein domains is not necessary for catalytic activity or substrate binding in a xylanase. *Biochem. J.* 269, 261-264

Ferreira, L. M., Wood, T. M., Williamson, G., Faulds, C., Hazlewood, G. P., Black, G. W. & Gilbert, H. J. (1993) A modular esterase from Pseudomonas fluorescens subsp. cellulosa contains a non-catalytic cellulose-binding domain. *Biochem. J.* 294, 349–355.

Ficko-Blean, E. and Boraston, A. B. (2005) Cloning, recombinant production, crystallization and preliminary X-ray diffraction studies of a family 84 glycoside hydrolase from Clostridium perfringens. *Acta Crystallogr Sect F Struct Biol Cryst Commun.* 61, 834-836

Flint, J., Nurizzo, D., Harding, S. E., Longman, E., Davies, G. J., Gilbert, H. J., and Bolam, D. N. (2004) Ligand-mediated dimerization of a carbohydrate-binding molecule reveals a novel mechanism for protein-carbohydrate recognition. *J. Mol. Biol*. 337, 417–426

Fontaine, T., Simenel, C., Dubreucq, G., Adam, O., Delepierre, M., Lemoine, J., Vorgias, C. E., Diaquin, M. and Latge, J. P. (2002) Molecular organization of the alkali-insoluble fraction of Aspergillus fumigatus cell wall. *J. Biol. Chem.* 275(36), 27594-607

Fontes, C. M. G. A., Clarke, J. H., Hazlewood, G. P., Fernandes, T. H., Gilbert, H. J. and Ferreira, L. M. A. (1998) Identification of tandemly repeated type VI cellulose-binding domains in an endoglucanase from the aerobic soil bacterium *Cellvibrio mixtus*. *Appl. Microbiol. Biotechnol*. 49, 552-559.

Freelove, A. C., Bolam, D. N., White, P., Hazlewood, G. P. and Gilbert, H. J. (2001) A novel carbohydrate binding protein is a component of the plant cell walldegrading complex of Piromyces equi. *J. Biol. Chem*. 276, 43010–43017.

Fry, S. C. (1989) Cellulases, hemicelluloses and auxin-stimulated growth: a possible relationship. *Physiol. Plant*. 75, 532-536.

Garcia-Martinez, D. V., Shinmyo, A., Madia, A. and Demain, A. L. (1980) Studies on cellulose production by *Clostridium thermocellum*. *Eur. J. Appl. Microbiol. Biotechnol.* 9, 189-197.

Gilbert, H. J., Bolam, D. N., Szabo, L., Xie,, H., Williamson, M. P., Simpson, P. J., Jamal, S., Boraston, A. B., Kilburn, D. G., Anthony, R., and Warren, J. (2002) An update in Carbohydrate-Binding Modules. Recent Advances in Carbohydrate Bioengineering (Teeri, T. T., Gilbert, H. J. and Feizi, T., Eds), pp. 89-98, *The Royal Society of Chemistry*, Cambridge, UK.

Gilkes, N. R., Kilburn, D. G., Warren, R. A. J. and Miller, Jr, R. C. (1988) Precise excision of the cellulose binding domains form two Cellulomonas fimi cellulases by a homologous protease and the effect on catalysis. *J. Biol. Chem.* 263, 10401–10407

Gloster, T. M., Ibatullin, F. M., Macauley, K., Eklof, J. M., Roberts, S., Turkenburg, J. P., Bjornvad, M. E., Jorgensen, P. L., Danielsen, S., Johansen, K. S., Borchert, T. V., Wilson, K. S., Brumer, H. and Davies, G. J. (2007) Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. *J Biol Chem.* 282, 19177-19189

Ha, M. A., Apperley, D. C., Jarvis, M.C. (1997) Molecular Rigidity in Dry and Hydrated Onion Cell Walls. *Plant Physiol.* 115(2): 593-598.

Hall, J., Black, G. W., Ferreira, L. M., Millward-Sadler, S. J., Ali, B. R., Hazlewood, G. P. and Gilbert, H. J. (1995) The non-catalytic cellulose-binding domain of a novel cellulase from Pseudomonas fluorescens subsp. cellulosa is important for the efficient hydrolysis of Avicel. *Biochem. J*. 309, 749–756

Hashimoto, H. (2006) Recent structural studies of carbohydrate-binding modules. *Cell. Mol. Life Sci*. 63, 2954–2967.

Hatada, Y., Hidaka, Y., Nogi, Y., Uchimura, K., Katayama, K., Li, Z., Akita, M., Ohta, Y., Goda, S., Ito, H., Matsui, H., Ito, S. and Horikoshi, K. (2004) Hyper-production of an isomalto-dextranase of an Arthrobacter sp. by a proteases-deficient Bacillus subtilis: sequencing, properties, and crystallization of the recombinant enzyme. *Appl Microbiol Biotechnol*. 65, 583-592.

Hayashi, T. (1989) Xyloglucans in the primary cell wall. *Annu. Rev. Plant Physiol. Plant Mol. Biol*. 40, 139-168

Hazlewood GP, Laurie JI, Ferreira LM, Gilbert HJ (1992) *Pseudomonas fluorescens* subsp. *cellulosa*: An alternative model for bacterial cellulose. *J Appl Bacteriol* 72:244–251.

Henshaw, J. L., D. N. Bolam, et al. (2004) The Family 6 Carbohydrate Binding Module CmCBM6-2 Contains Two Ligand-binding Sites with Distinct Specificities. *J. Biol. Chem.* 279 (20): 21552-21559.

Henshaw, J., Horne, A., van Bueren, A. L., Money, V. A., Bolam, D. N., Czjzek, M., Ekborg, N. A., Weiner, R. M., Hutcheson, S. W., Davies, G. J., Boraston, A. B., Gilbert, H. J. (2006) Family 6 carbohydrate binding modules in beta-agarases display exquisite selectivity for the non-reducing termini of agarose-chains. *J. Biol. Chem.* 281, 17099-17107.

Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J*. 280, 309–316.

Henrissat, B. (1998) Glycosidase families. *Biochem Soc Trans* 26(2): 153-6.

Henrissat, B., Coutinho, P. M., Davies, G. J. (2001) A census of carbohydrate-active enzymes in the genome of. *Arabidopsis thaliana*. *Plant Mol Biol* 47, 55–72.

Hermoso, J. A., Sanz-Aparicio, J., Molina, R., Juge, N., Gonzalez, R., Faulds, C. B. (2004) The crystal structure of feruloyl esterase a from Aspergillus niger suggests evolutive functional convergence in feruloyl esterase family. *J Mol Biol*, 338, 495-506.

Ho, Y. S., Swenson, L., Derewenda, U., Serre, L., Wei, Y., Dauter, Z. et al. (1997) Brain acetylhydrolase that inactivates platelet-activating factor is a G-protein-like trimer. *Nature*, 385, 89–93.

Hogg, D., Pell, G., Dupree, P., Goubet, F., Martin-Orue, S. M., Armand, S. and Gilbert, H. J. (2003) The modular architecture of Cellvibrio japonicus mannanases in glycoside hydrolase families 5 and 26 points to differences in their role in mannan degradation. *Biochem J.* 371(Pt 3): 1027-43.

Jamal, S., Nurizzo, D., Boraston, A. B. and Davies, G. J. (2004) X-ray crystal structure of a non-crystalline cellulose-specific carbohydrate-binding module: CBM28. *J Mol Biol.* 339(2): 253-258.

Jamal-Talabani, S., Boraston, A. B., Turkenburg, J. P., Tarbouriech, N., Ducros, V. M. and Davies, G. J. (2004) Ab initio structure determination and functional characterization of CBM36; a new family of calcium-dependent carbohydrate binding modules. *Structure* 12, 1177–1187

Jenkins, J., Mayans, O., Smith, D., Worboys, K., Pickersgill, R. W. (2001) Threedimensional structure of Erwinia chrysanthemi pectin methyl esterase reveals a novel esterase active site. *J Mol Biol*, 305, 951-960.

Jervis, E. J., Haynes, C. A. and Kilburn, D. G. (1997) Surface diffusion of cellulases and their isolated binding domains on cellulose. *J Biol Chem*. 272, 24016-24023

Johnson, K. G., Fontana, J. D. & MacKenzie, C. R. (1988) Measurement of acetylxylan esterase in Streptomyces. *Method. Enzymol*. 160, 551–560.

Jones, S. and Thornton, J. M. (1995) Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* 63, 31–65.

Kauss, H., Hassid, W.Z. (1967) Biosynthesis of the 4-O-methyl-D-glucuronic acid unit of hemicellulose B by transmethylation from S-adenosyl-L-methionine. *J Biol Chem* 242, 1680–1684.

Keegstra, K., Talmadge, K. W., Bauer, W. D. and Albersheim, P. (1973) The structure of plant cell walls III. A model of the walls of suspension-cultured sycamore cells based on the interconnections of the macromolecular components. *Plant Physiol.* 51, 188-196

Kellett, L. E., D. M. Poole, et al. (1990) Xylanase B and an arabinofuranosidase from Pseudomonas fluorescens subsp. cellulosa contain identical cellulose-binding domains and are encoded by adjacent genes. *Biochem J* 272(2), 369-76.

Kloareg, B. and Quatrano, R. S. (1988) Structure of the cell walls of marine algae and ecophysiological functions of the matrix of polysaccharides. *Oceanogr Mar Biol Annu Rev* 26, 259–315.

Knudsen, K. E. B. (1997) Carbohydrate and lignin contents of plant materials used in animal feeding. *Anm. Feed Sci. Technol.* 67, 319-338.

Kofod, L. V., Kauppinen, S., Christgau, S., Andersen, L. N., Heldt-Hansen, H. P., Dorreich, K. and Dalboge, H. (1994) Cloning and characterization of two structurally and functionally divergent rhamnogalacturonases from Aspergillus aculeatus. *J Biol Chem.* 269(46): 29182-29189.

Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr*. D, 60, 2256–2268.

Kroon, P. A., Williamson, G., Fish, N. M., Archer, D. B. and Belshaw, N. J. (2000) A modular esterase from Penicillium funiculosum which releases ferulic acid from plant cell walls and binds crystalline cellulose contains a carbohydrate binding module. *Eur J Biochem.* . 267, 13

Lamed, R. and Zeikus, J. G. (1980) Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in Clostridium thermocellum and Thermoanaerobium brockii. *J Bacteriol* 144(2): 569-78.

Lamzin, V. S. and Wilson, K. S. (1997) Automated refinement for protein crystallography. *Methods Enzymol*. 277, 269–305.

Laskowski, R. A., Moss, D. S. and Thornton, J. M. (1993) Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol*. 231, 1049–1067.

Lehtio, J., Sugiyama, J., Gustavsson, M., Fransson, L., Linder, M. and Teeri, T. T. (2003) The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules. *Proc Natl Acad Sci U S A* 100(2): 484-9.

Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* 26.

Levy, I. and Shoseyov, O. (2002) Cellulose-binding domains: biotechnological applications. *Biotechnol Adv* 20(3-4): 191-213.

Lo, Y. C., Lin, S. C., Shaw, J. F. & Liaw, Y. C. (2003) Crystal structure of Escherichia coli thioesterase I/protease I/lysophospholipase L1: consensus sequence blocks constitute the catalytic center of SGNHhydrolases through a conserved hydrogen bond network. *J. Mol. Biol*. 330, 539–551.

Lo, Y. C., Lin, S. C., Shaw, J. F. & Liaw, Y. C. (2005) Substrate specificities of Escherichia coli thioesterase I/protease I/lysophospholipase L1 are governed by its switch loop movement. *Biochemistry*, 44, 1971–1979.

Lymar, E. S., Li, B. and Renganathan, V. (1995) Purification and Characterization of a Cellulose-Binding (beta)-Glucosidase from Cellulose-Degrading Cultures of Phanerochaete chrysosporium. *Applied and Environmental Microbiology*. 61, 5

Lytle, B. L., Volkman, B. F., Westler, W. M., Heckman, M. P. and Wu, J. H. (2001) Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain. *J Mol Biol.* 307, 745-753

Lytle, B. L., Volkman, B. F., Westler, W. M. and Wu, J. H. (2000) Secondary structure and calcium-induced folding of the Clostridium thermocellum dockerin domain determined by NMR spectroscopy. *Arch Biochem Biophys.* 379, 237-244

Marcus, S. E., Y. Verhertbruggen, et al. (2008) Pectic homogalacturonan masks abundant sets of xyloglucan epitopes in plant cell walls. *BMC Plant Biol* 8, 60.

Maurice, S., Dekel, M., Shoseyov, O. and Gertler, A. (2003) Cellulose beads bound to cellulose binding domain-fused recombinant proteins; an adjuvant system for parenteral vaccination of fish. *Vaccine* 21(23): 3200-7.

Maynes, J. T., Garen, C., Cherney, M. M., Newton, G., Arad, D., Av-Gay, Y., Fahey, R. C., James, M. N. G. (2003) The crystal structure of 1-D-myoinosityl-2-acetamido-2-deoxy-alpha-D-glucopyranoside deacetylase (MshB) from Mycobacterium tuberculosis reveals a zinc hydrolase with a lactate dehydrogenase fold. *J Biol Chem*, 278, 47166-47170.

McCartney, L., H. J. Gilbert, D. N. Bolam, A. B. Boraston, and J. P. Knox (2004) Glycoside hydrolase carbohydrate-binding modules as molecular probes for the analysis of plant cell wall polymers. *Anal. Biochem*. 326, 49–54.

McCartney, L., Blake, A. W., Flint, J., Bolam, D. N., Boraston, A. B., Gilbert, H. J. and Knox, J. P. (2006) Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules. *Proc. Nat. Acad. Sci. USA*, 103, 4765-4770

McDonald, I. K. and Thornton, J. M. (1994) Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238, 777–793.

McGreal, E. P., L. Martinez-Pomares, et al. (2004) Divergent roles for C-type lectins expressed by cells of the innate immune system. Molecular Immunology 41, 1109-1121.

Miller, G. L. (1959) The use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal. Chem*. 31, 426-428.

Miyanaga, A., Koseki, T., Matsuzawa, H., Wakagi, T., Shoun, H. and Fushinobu, S. (2004) Crystal structure of a family 54 alpha-L-arabinofuranosidase reveals a novel carbohydrate-binding module that can bind arabinose. *J. Biol. Chem.* 279(43), 44907-44914

Molgaard, A., Kauppinen, S. and Larsen, S. (2000) Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases. Structure, 8, 373–383.

Molgaard, A. and Larsen, S. (2002) A branched N-linked glycan at atomic resolution in the 1.12 Å structure of rhamnogalacturonan acetylesterase. *Acta Crystallog. sect. D*, 58: 111–119.

Montanier, C., van Bueren, A. L., Dumon, C., Flint, J. E., Correia, M. A., Prates, J. A., Firbank, S. J., Lewis, R. J., Grondin, G. G., Ghinet, M. G., Gloster, T. M., Herve, C., Knox, J. P., Talbot, B. G., Turkenburg, J. P., Kerovuo, J., Brzezinski, R., Fontes, C. M., Davies, G. J., Boraston, A. B. and Gilbert, H. J. (2009) Evidence that family 35 carbohydrate binding modules display conserved specificity but divergent function. *Proc Natl Acad Sci U S A.* 106, 3065-3070

Moran, F., Nasuno, S. and Starr, M.P. (1968) Extracellular and intracellular polygalacturonic acid trans-eliminases of *Erwinia carotovora*. *Arch Biochem Biophys* 123, 298 –306.

Murshudov, G. N., Vagin, A. A. and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. Biol. D, 53, 240–255.

Nagy, T., Simpson, P., Williamson, M. P., Hazlewood, G. P., Gilbert, H. J. and Orosz, L. (1998) All three surface tryptophans in Type IIa cellulose binding domains play a pivotal role in binding both soluble and insoluble ligands. *FEBS Lett.* 429, 312-316

Najmudin, S., Guerreiro, C. I., Carvalho, A. L. Prates, J. A., Correia, M. A., Alves, V. D., Ferreira, L. M., Romao, M. J., Gilbert, H. J., Bolam, D. N. and Fontes, C. M. (2006) Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *J. Biol. Chem*. 281, 8815-8828.

Navarro-Fernandez, J., Martinez-Martinez, I., Montoro-Garcia, S., Garcia-Carmona, F., Takami, H. and Sanchez-Ferrer, A. (2008) Characterization of a new rhamnogalacturonan acetyl esterase from Bacillus halodurans C-125 with a new putative carbohydrate binding domain. *J Bacteriol.* 190(4): 1375-1382.

Navaza, J. (1994) *AMoRe*: An automated package for molecular replacement. *Acta Crystallogr D* 50:157–163.

Nisizawa, K., Yamaguchi, T., Handa, N., Maeda, M., Yamazaki, H. (1963). Chemical Nature of a Uronic Acid-Containing Polysaccharide in the Peritrophic Membrane. J Biochem 54, 419-426

Notenboom, V., Boraston, A. B., Chiu, P., Freelove, A. C., Kilburn, D. G., and Rose, D. R. (2001) Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: an X-ray crystallographic, thermodynamic and mutagenic study. *J. Mol. Biol.* 314, 797-806

Notenboom, V., Boraston, A. B., Williams, S. J., Kilburn, D. G. and Rose, D. R. (2002) High-resolution crystal structures of the lectin-like xylan binding domain from Streptomyces lividans xylanase 10A with bound substrates reveal a novel mode of xylan binding. *Biochemistry* 41(13): 4246-54.

O'Sullivan, A. C. (1997) Cellulose: the structure slowly unravels. *Cellulose.* 4: 173-207.

Otwinowski Z and Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276:307–326.

Pala, H., Lemos, M.A., Mota, M., and Gama, F.M. (2001) Enzymatic upgrade of old paperboard containers. *Enzyme microb. Technol.* 29, 274-279.

Pape, T. and Schneider, T. R. (2004) HKL2MAP: a graphical user interface for phasing with SHELX programs. J. Appl. Crystallogr. 37, 843–844.

Pena, M. J., Zhong, R., Zhou, G. K., Richardson, E. A., O'Neill, M. A., Darvill, A. G., York, W. S. and Ye, Z. H. (2007) Arabidopsis irregular xylem8 and irregular xylem9: implications for the complexity of glucuronoxylan biosynthesis. *Plant Cell* 19(2), 549-63.

Pell, G., Williamson, M. P., Walters, C., Du, H., Gilbert, H. J. and Bolam, D. N. (2003) Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from Cellvibrio japonicus Xyn10C. *Biochemistry.* 42, 9316-9323

Pflugrath J. W. (1999) The finer things in X-ray diffraction data collection. *Acta Crystallogr D* 55:1718–1725.

Pires, V. M. R., Henshaw, J. L., Prates, J. A. M., Bolam, D. N., Ferreira, L. M. A., Fontes, C. M. G. A., Henrissat, B., Planas, A., Gilbert, H. J. and Czjzek, M. (2004) The Crystal Structure of the Family 6 Carbohydrate Binding Module from Cellvibrio mixtus Endoglucanase 5A in Complex with Oligosaccharides Reveals Two Distinct Binding Sites with Different Ligand Specificities. *The Journal of Biological chemistry*. 279 (20), 21560-21568

Prates, J. A., Tarbouriech, N., Charnock, S. J., Fontes, C. M., Ferreira, L. M. and Davies, G. J. (2001) The structure of the feruloyl esterase module of xylanase 10B from Clostridium thermocellum provides insights into substrate recognition. *Structure*. 9, 1183-1190

Proctor, M. R., Taylor, E. J., Nurizzo, D., Turkenburg, J. P., Lloyd, R. M., Vardakou, M., Davies, G. J., and Gilbert, H. J. (2005) *Proc Nat Acad Sci USA*, 102, 2697-2702.

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A. et al. (2006) The path forward for biofuels and biomaterials. Science, 311, 484–489.

Raghothama, S., Simpson, P. J., Szabo, L., Nagy, T., Gilbert, H. J. and Williamson, M. P. (2000) Solution structure of the CBM10 cellulose binding module from *Pseudomonas* xylanase A. *Biochemistry.* 39, 978-984

Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. Nature, 362, 219–223.

Read, S. M., Currie, G. and Bacic, A. (1996) Analysis of the structural heterogeneity of laminarin by electrospray-ionisation-mass spectrometry. *Carbohydr Res* 281, 187-201.

Richins, R. D., Mulchandani, A. and Chen, W. (2000) Expression, immobilization, and enzymatic characterization of cellulose-binding domain-organophosphorus hydrolase fusion enzymes. *Biotechnol Bioeng* 69(6): 591-6.

Sakai K, Katsumi R, Isobe A, Nanjo F (1991) Purification and hydrolytic action of a chitosanase from Nocardia orientalis. *Biochim Biophys Acta* 1079:65–72.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor laboratory (2nd Ed.), Cold Spring Harbor, New York.

Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with SHELXD. *Acta Crystallogr.D*, 58, 1772–1779.

Simpson, P. J., H. Xie, et al. (2000) The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J Biol Chem.* 275(52): 41137-42.

Schell, D. J., Riley, C. J., Dowe. N., Farmer, J., Ibsen, K. N., Ruth, M. F., Toon, S. T. and Lumpkin, R. E. (2004) A bioethanol process development unit: initial operating experiences and results with a corn fiber feedstock. *Bioresour. Technol*. 91, 179-88.

Shoseyov, O., Shani, Z., Levy, I. (2006) Carbohydrate binding modules: biochemical properties and novel applications. Microbiol Mol Biol Rev., 70(2), 283-95

Solbak, A. I., T. H. Richardson, et al. (2005) Discovery of pectin-degrading enzymes and directed evolution of a novel pectate lyase for processing cotton fabric. *J Biol Chem* 280(10), 9431-8.

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., Osborne, E., Paredez, A., Persson, S., Raab, T., Vorwerk, S. and Youngs, H. (2004) Toward a Systems Approach to Understanding Plant Cell Walls. *Science*. 306, 6

Stålbrand H., Saloheimo A., Vehmaanperä J., Henrissat B, Penttilä M. (1995) Cloning and expression in Saccharomyces cerevisiae of a Trichoderma reesei beta-mannanase gene containing a cellulose binding domain. *Appl. Environ. Microbiol. Appl Environ Microbiol*. 1995 Mar; 61(3):1090-7

Sticklen, M. (2006) Plant genetic engineering to improve biomass characteristics for biofuels. *Curr. Opin. Biotechnol*. 17, 315–319.

Suurnäkki, A., Tenkanen, M., Siika-Aho, M., Niku-Paavola, M-L., Viikari, L. and Buchert, J. (2000) Structure of a family 15 carbohydrate-binding module in complex with xylopentaose.Evidence that xylans binds in an approximate 3-fold helical conformation. *J. Biol. Chem.* 276, 49061-65.

Talbott, L. D. and P. M. Ray (1992) Molecular Size and Separability Features of Pea Cell Wall Polysaccharides : Implications for Models of Primary Wall Structure. *Plant Physiol* 98(1), 357-368.

Takeo, K. (1984) Affinity electrophoresis - principles and applications. *Electrophoresis* 5, 187-195

Taylor, E. J., Gloster, T. M., Turkenburg, J. P., Vincent, F., Brzozowski, A. M., Dupont, C. et al. (2006) Structure and activity of two metal ion-dependent acetylxylan esterases involved in plant cell wall degradation reveals a close similarity to peptidoglycan deacetylases. *J. Biol. Chem*. 281, 10968–10975.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22, 4673–4680.

Tomme, P., Van Tilbeurgh, H., Pettersson, G., Van Damme, J., Vandekerckhove, J., Knowles, J., Teeri, T. and Claeyssens, M. (1988) Studies of the cellulolytic system of Trichoderma reesei QM 9414: analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur. J. Biochem*. 170, 575–581

Tomme, P., Warren, R. A. and Gilkes, N. R. (1995) Cellulose hydrolysis by bacteria and fungi. *Adv Microb Physiol* 37: 1-81.

Tomme, P., Boraston, A., Kormos, J. M., Warren, A. J. and Kilburn, D. G. (2000) Affinity electrophoresis for the identification and characterization of soluble sugar binding by carbohydrate-binding modules. *Enz. Microb. Technol*. 27, 453-458.

Tormo, J., Lamed, R., J.Chirinol, A., Morag, E., A.Bayer, E., Shoham, Y. and A.Steitz, T. (1996) Crystal structure of a bacterial family-III cellulose- binding domain: a general mechanism for attachment to cellulose. *EMBO Journal*.15(21), 5739-51.

Tunnicliffe, R.B., Bolam, D.N., Pell, G., Gilbert, H.J., Williamson, M.P. (2005). Structure of a mannan-specific family 35 carbohydrate-binding module: evidence for significant conformational changes upon ligand binding. *J Mol Biol*. 347(2), 287-96.

Vagin, A. and Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.* 30, 1022–1025.

Vagin A, Teplyakov A (2000) An approach to multi-copy search in molecular replacement. *Acta Crystallogr D* 56:1622–1624.

Van Bueren, A. L., Morland, C., Gilbert, H. J. and Boraston, A. B. (2005) Family 6 carbohydrate binding modules recognize the non-reducing end of beta-1,3-linked glucans by presenting a unique ligand binding surface. *J. Biol Chem.* 280, 530-537.

Van Bueren, A. L., Higgins, M., Wang, D., Burke, R. D. and Boraston, A.B. (2007) Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors. *Nat Struct Mol Biol* 14:76–84.

Verma, D. P. and Hong, Z. (2001) Plant callose synthase complexes. *Plant Mol Biol* 47, 693-701.

Vijayan, M. and Chandra, N. (1999) Lectins. *Curr Opin Struct Biol.* 9, 707-714

Vincent, F., Charnock, S. J., Verschueren, K. H., Turkenburg, J. P., Scott, D. J., Offen, W. A. et al. (2003) Multifunctional xylooligosaccharide/cephalosporin C deacetylase revealed by the hexameric structure of the Bacillus subtilis enzyme at 1.9 Å resolution. *J. Mol. Biol*. 330, 593–606.

Vincent, F., Yates, D., Garman, E., Davies, G. J., Brannigan, J. A. (2004) The 3-D structure of the N-acetylglucosamine-6-phosphate deacetylase, NagA, from Bacillus subtilis: a member of the urease superfamily. *J Biol Chem*, 279, 2809-2816.

Wang, A. A., A. Mulchandani, et al. (2002) Specific adhesion to cellulose and hydrolysis of organophosphate nerve agents by a genetically engineered Escherichia coli strain with a surface-expressed cellulose-binding domain and organophosphorus hydrolase. *Appl Environ Microbiol* 68(4): 1684-9.

Wang, J., Stuckey, J. A., Wishart, M. J. and Dixon, J. E. (2002) A unique carbohydrate binding domain targets the lafora disease phosphatase to glycogen. *J Biol Chem* 277:2377–2380.

Whittington, D. A., Rusche, K. M., Shin, H., Fierke, C. A. and Chistianson DW (2003) Crystal structure of LpxC, a zinc-dependent deacetylase essential for endotoxin biosynthesis. *Proc Natl Acad Sci USA*, 100, 8146-8150.

Xie, H., Bolam, D. N., Nagy, T., Szabo, L., Cooper, A., Simpson, P. J., Lakey, J. H., Williamson, M. P., and Gilbert, H. J. (2001) Role of hydrogen bonding in the interaction between a xylan binding module and xylan. *Biochemistry* 40, 5700-5707

Xu, Z., Bae, W., Mulchandani, A., Mehra, R. K. and Chen, W. (2002) Heavy metal removal by novel CBD-EC20 sorbents immobilized on cellulose. *Biomacromolecules* 3(3): 462-5.

Xu, Q., Bayer, E. A., Goldman, M., Kenig, R., Shoham, Y. and Lamed, R. (2004) Architecture of the Bacteroides cellulosolvens cellulosome: description of a cell surface-anchoring scaffoldin and a family 48 cellulase. *J Bacteriol* 186(4): 968-77.

Zverlov, V. V., Volkov, I. Y., Velikodvorskaya, G. A. and Schwarz, W. H. (2001) The binding pattern of two carbohydrate-binding modules of laminarinase Lam16A from Thermotoga neapolitana : differences in *β*-glucan binding within family CBM4. *Microbiology* 147, 621–629